electronics

MDPI

Article

SVM-Based Blood Exam Classification for Predicting Defining Factors in Metabolic Syndrome Diagnosis

Dimitrios P. Panagoulias, Dionisios N. Sotiropoulos and George A. Tsihrintzis *

Department of Informatics, University of Piraeus, Karaoli and Dimitriou 80, 185 34 Piraeus, Greece; panagoulias_d@yahoo.gr (D.P.P.); dsotirop@unipi.gr (D.N.S.)
* Correspondence: geoatsi@unipi.gr

Abstract: Biomarkers have already been proposed as powerful classification features for use in the training of neural network-based and other machine learning and artificial intelligence-based prognostic models in the scientific field of personalized nutrition. In this paper, we construct and study cascaded SVM-based classifiers for automated metabolic syndrome diagnosis. Specifically, using blood exams, we achieve an average accuracy of about 84% in correctly classifying body mass index. Similarly, cascaded SVM-based classifiers achieve a 74% accuracy in correctly classifying systolic blood pressure. Next, we propose and implement a system that achieves an 84% accuracy in metabolic syndrome prediction. The proposed system relies not only on prediction of the body mass index but also on prediction from blood exams of total cholesterol, triglycerides and glucose. For the aim of self-completeness of the paper, the key concepts with regard to metabolic syndrome are summarized, and a review of previous related work is included. Finally, conclusions are drawn and indications for related future research are outlined.

Keywords: biomarkers; metabolic syndrome (MetS); body mass index (BMI); systolic blood pressure (SBP); personalized medicine; support vector machine (SVM); support vector classifier (SVC); neural network; data analytics

Citation: Panagoulias, D.P.; Sotiropoulos, D.N.; Tsihrintzis, G.A. SVM-Based Blood Exam Classification for Predicting Defining Factors in Metabolic Syndrome Diagnosis. *Electronics* **2022**, *11*, 857. https://doi.org/10.3390/electronics11060857

Academic Editors: Juan M. Corchado, In Lee, Fuji Ren, Rashid Mehmood, Byung-Gyu Kim and Carlos A. Iglesias

Received: 2 February 2022
Accepted: 4 March 2022
Published: 9 March 2022

1. Introduction

In the recent years, there is a growing interest in the incorporation of artificial intelligence technologies—including machine learning and deep learning—in healthcare and medicine. These technologies are expected to have a transformative role in patient treatment and disease management via automating tasks, streamlining processes, keeping manual intervention at a minimum and simplifying mundane operations for all parties involved [1].

On the other hand, biomarkers are objective measurements drawn from blood and other bodily fluids or tissue that form medical signs or indicators of a disease or, more generally, the health state of a person. Biomarkers can comprise either a sole metric or a combination of metrics and observations [2]. In Table 1, some commonly used biomarkers are presented, including waist-to-hip ratio, total cholesterol, systolic blood pressure (SBP) and fasting glucose. Currently, biomarkers are seen as key drivers of the personalization of patient management and treatment and drug development.

In this paper, we report on recent findings from our research on investigating the link between a person's standard biochemistry profile (based on blood exams), his/her body mass index (BMI), metabolism as health state and SBP. Our current findings expand upon our previous related research, which was based on the use of deep neural networks and other machine learning paradigms in relation to BMI and nutrition [3–6].

The motivation for this research was to compress tasks and improve outcomes by using more common variables to predict health states and, in a sense, simplify the patient's journey via minimizing response time between test, result and recommended action. At the

same time, another motivation was to find methods to optimize operational issues via applying machine learning methodologies that can be easily transcribed into telemedicine applications. For example, Big Data and artificial intelligence can be used to improve decision making, support interventions [7] and add more pathways in healthcare analytics [8]. Mathematical tools and Big Data, as part of advanced machine learning pipelines and artificial intelligence, will eventually become the basis for analysis in diagnostics and pathology [9].

Table 1. Examples of biomarkers.

Name	Description	Health Results
Waist to hip ratio	Abdominal obesity index	Hypertension, CHD, insulin-dependent diabetes and stroke
Total cholesterol	Helps in the synthesis of bile acids and steroid hormones	At middle age: Coronary heart disease (CHD) and mortality of all causes at an older age: u-shaped relation to death
Systolic blood pressure (SBP)	Cardiovascular activity index: maximum pressure in an artery when the heart supplies the body with blood	Cardiovascular death (CVD), stroke, coronary heart disease (CHD)
Fasting glucose	Measures the amount of sugar in the diabetes index	Diabetes, CHD, mortality, poor cognitive function

Childhood obesity is a very important issue and is connected to genetic predispositions and behavior. According to the CDC and based on recent studies, obesity prevalence rises to approximately 13% in the age group of 2 to 5 years, 20% in the age group 6 to 11 years and 21.2% among 12 to 19 year olds [10]. Furthermore, childhood obesity is commonly associated with certain communities and specific populations [10]. The age ranges of our dataset can be seen in Figure 1. Clearly, mostly young adults and people above the age of 18 are included, with only a small number of them belonging to the 7–14 age group. Thus, additional datasets need to be collected to draw reliable conclusions with regard to childhood obesity. Demographics and lifestyle determinants could also prove to be useful tools if used as inputs, alongside other nutritional factors and laboratory data, for machine learning classifiers and should be investigated further. Machine learning tools could be used as predictors of child obesity or deployed for automating targeted interventions. This research avenue is, in fact, being followed, and its results will be announced in other fora.

In this study, we are looking into the spectrum of metabolic syndrome (MetS) as seen in Figure 2. We follow a more holistic method by looking into engineering shortcuts using linked automation. Via building on related literature [11] and expanding on related conclusions [12], we propose a more complete approach via linking factors to states and via using said states to extract actions. MetS being a specific health state and weight being a factor related to MetS and triglyceride combined with glucose and cholesterol are accompanied by defining factors; we utilize all three to finalize a conclusion and calculate related risk factors and recommendations. In this paper, we implement and test various classifiers towards linking a person's standard biochemistry profile (based on blood exams), his/her BMI, metabolism as health state and SBP. We show that support vector machine-based classifiers are very promising. We also provide a brief comparison with results on previous works of ours that were based on deep neural networks [3–6]. Moreover, a more extensive look into related literature is provided in Section 3.

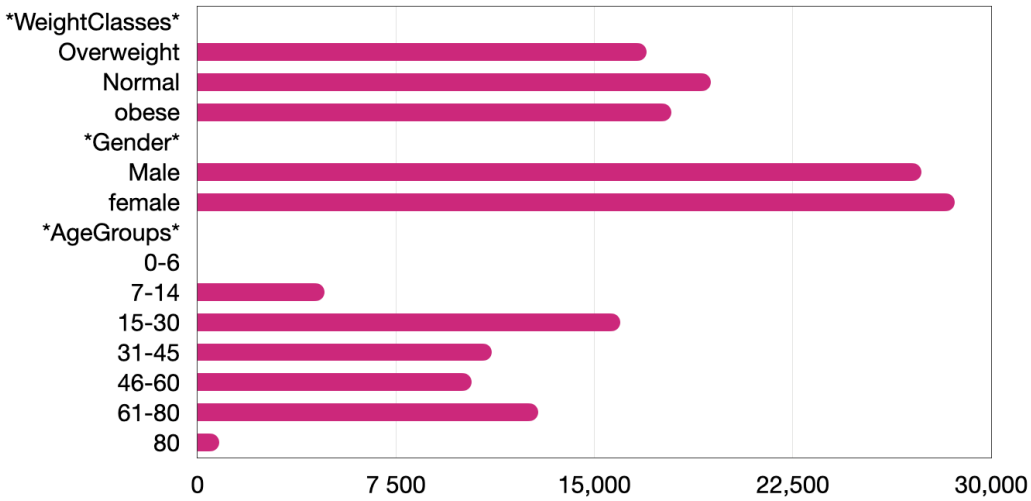

Figure 1. Characteristics of sample.

More specifically, the paper is organized as follows: In Section 2, the key concepts with regard to MetS are summarized, laying the theoretical ground work on which our research is based. In Section 3, previous related work is highlighted. In Section 4, we discuss the datasets used in our research, as well as important statistical measures to characterize them. In Section 5, we develop and comparatively evaluate various classifiers for BMI prediction based on blood exams, including neural network-based and SVM-based classifiers. Furthermore, we show that a cascaded SVM-based classifier is most promising, achieving an average correct classification rate of about 84%. In Section 6, we propose and implement a system for MetS prediction, which relies not only on BMI prediction but also on the prediction from blood exams of all MetS defining factors, i.e., total cholesterol, triglycerides and blood pressure (Figure 2). In Section 7, we itemize the key findings of our research and discuss their significance. Finally, in Section 8, we draw conclusions and point to future research avenues in this area.

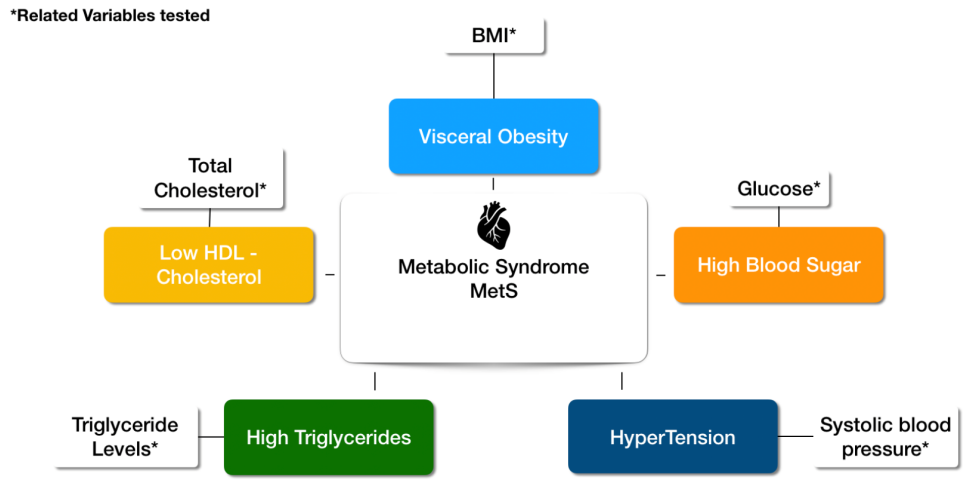

Figure 2. MetS Predictors.

2. Metabolic Syndrome

Noncommunicable diseases (NCDs) are considered a major cause of morbidity and mortality in the developed world but more significantly so in underdeveloped countries. MetS has been prominent among the NCDs. Specifically, MetS, also known as syndrome X or dysmetabolic syndrome, refers to a cluster of metabolic conditions that can lead, among other things, to heart disease. The main features of MetS include insulin resistance, hypertension (i.e., high blood pressure), abnormal cholesterol and an increased risk for

clotting. The results of most studies have documented overweight and obesity as the strongest predictors of MetS. On the other hand, thyroid dysfunction is also one of the most common endocrine disorders in MetS patients [13].

2.1. Metabolic Syndrome Markers

MetS is one of the many risk factors for atherosclerotic cardiovascular disease (ASCVD) and, thus, represents a combination of determinants of risk that compares to that of cigarette smoking, hypertension, hypercholesterolemia and diabetes [14]. Atherogenic dyslipidemia, high blood pressure, high glucose, prothrombotic condition and pro-inflammatory condition are components frequently found in MetS. Clinical hyperglycemia (Type 2 diabetes) is also observed in more advanced stages of MetS. MetS is labeled as a strong cardiovascular risk feature and, even without diabetes being present, it doubles the relative risk for cardiovascular disease. MetS is also considered a worldwide problem and, as urbanisation leads to obesity expansion among the population due to lifestyle shifts, MetS becomes proportionally more frequent. In the latest studies, the presence of MetS has also been associated with higher risk of acute respiratory distress syndrome (ARDS) and increase in deaths for patients with COVID-19 [15]. Currently, a medical rule of thumb associates metabolic disorders (MetS) with at least three of the conditions (biomarkers) mentioned in Figure 2, upon which we further elaborate in the following.

2.1.1. A Large Waist

BMI is more commonly used than waist circumference as a measure of adiposity in clinical and research settings [16]. Multiple factors have been proven to have some relativity or to contribute to MetS. However, it is rather uncommon to detect MetS when excess body fat is absent. Thus, MetS is more ubiquitous where obesity increases. In obese individuals, excess adipose tissue releases a variety of factors that may be contributing to metabolic risk factors. Specifically, excessive release of non-esterified fatty acids predisposes the individual to accumulation of ectopic fat in the liver, muscles and visceral adipose tissue stores [17].

2.1.2. A High Triglyceride Level

This is determined by blood tests that are used to gauge the amount of triglycerides, which constitute a type of fat in the blood. Levels are high when more fat is consumed than the amount necessary for body functions.

2.1.3. Reduced HDL ("Good" Cholesterol)

In our body, we accumulate both good (HDL) and bad (LDL) cholesterol. Bad cholesterol accumulates in arteries, whereas good cholesterol scavenges and removes it, keeping bad cholesterol from building up in the arteries. Ranges of acceptable values of both HDL and LDL differ between men and women.

On the other hand, the total cholesterol score is calculated using the following equation: HDL level + LDL level + 20% of the triglyceride level. Total cholesterol equals the overall amount of cholesterol in the blood, including both *high density lipoprotein* (HDL cholesterol) and *low density lipoprotein* (LDL cholesterol). High total cholesterol levels represent increased ischemic stroke probability [18]. HDL decreases the likelihood of heart problems and, invertedly, LDL increases the risk of stroke.

2.1.4. Elevated Fasting Blood Sugar (Fasting Glucose)

Carbohydrates are received from the consumption of foods and, through them, glucose ("blood sugar") is produced. Glucose is required by the body as an energy provider. Insulin is necessary for glucose to travel to cells and for energy to be released. In the absence of insulin, glucose levels rise; thus, diabetes occurs because of the disability of the pancreas to produce sufficient the levels of insulin required for metabolic processes to occur [19].

2.1.5. Increased Blood Pressure

High blood pressure or hypertension is the condition where the force of blood against the artery walls is so strong that, in the mid- to long-term, it can be a main cause of major health problems and, more likely, heart disease. The peak (systolic) number is the pressure when the heart beats. The lowest (diastolic) number is the pressure when the heart is resting between pulses. Normal blood pressure is below 120/80 mm Hg. High blood pressure is a systolic pressure of 140 mm Hg or greater and/or a diastolic pressure of 90 mm Hg or greater, which remains at high levels over time [20]. Blood pressure ranges are defined in more detail in Figure 3.

Blood Pressure Estimator

Category	Systolic(mm HG-upper#)		Diastolic(mm HG-lower#)
Normal	Less than 120	and	Less than 80
Elevated	120-129	and	Less than 80
High Blood Pressure(Hypertension) Stage 1	130-139	or	80-89
High Blood Pressure(Hypertension) Stage 2	140 or higher	or	90 or higher
Hypertensive Crisis	Higher than 180	and /or	Higher than 120

Figure 3. Blood pressure estimator.

2.2. Genetics—Cardiac Effects of Obesity and MetS

The genome contains numerous markers associated with MetS. In fact, there are hundreds of markers in the genome that are associated with the biological traits of MetS. Each genetic component can exist at many levels, e.g., within adipose tissue, in insulin signaling pathways and as regulatory functions of the individual components of the syndrome. Each level exhibits its own genetic background. As such, no common genetic trait has been identified for MetS [21].

The observed association of obesity with hypertension prompted a body of work exploring causes and effects of obesity on the heart [22]. Chronic increases in body weight and adiposity can result in significant neurohormonal changes and adaptations in the cardiovascular system [23]. The ratio of total cholesterol over high-density lipoprotein cholesterol (HDL-C) and the ratio of low-density lipoprotein cholesterol (LDL-C) over high-density lipoprotein cholesterol (HDL-C) are commonly used to predict ischemic heart disease risk [24].

3. Related Work and Methodology Comparison

In our previous study [3], we tested and evaluated the ability of neural networks to classify people in classes based on their BMI and using a basic biochemical profile extracted via routine blood exams (Figure 4). The classification process was firstly realized in four classes, then in three and finally in two (2) and a general understanding of the relation was established. Specifically, four BMI classes are defined in the relevant literature, namely "obese", "overweight", "normal" and "underweight". At each stage, accuracy increased while classes were thinned into smaller groups, as in Figure 5. Thus, the idea of a cascaded classifying method was conceived in accordance with similar previous approaches with regard to the recommendation problem [25,26]. At the same time, it became apparent that

blood exams held sufficient information for a system to be tested in a greater variety of health states. Going forward, the data used would be thoroughly analyzed to ensure that no bias was intended in the classification system. The results of each classifier would be cross-examined by using comparison matrices. In the current paper, we report on our study of binary and one-class support vector machine (SVM)-based classifiers, which we also compare with several other classifiers.

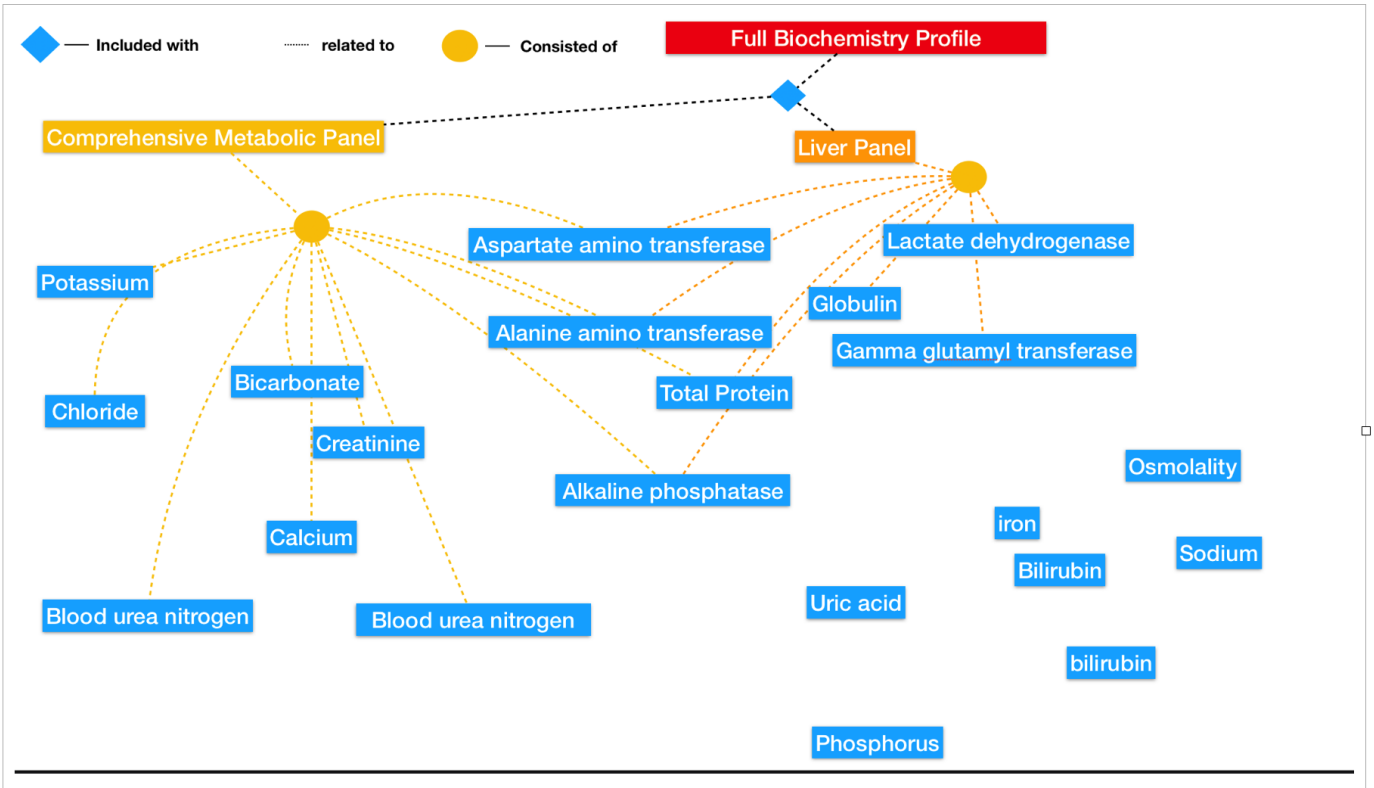

Figure 4. Full biochemistry profile—blood exams.

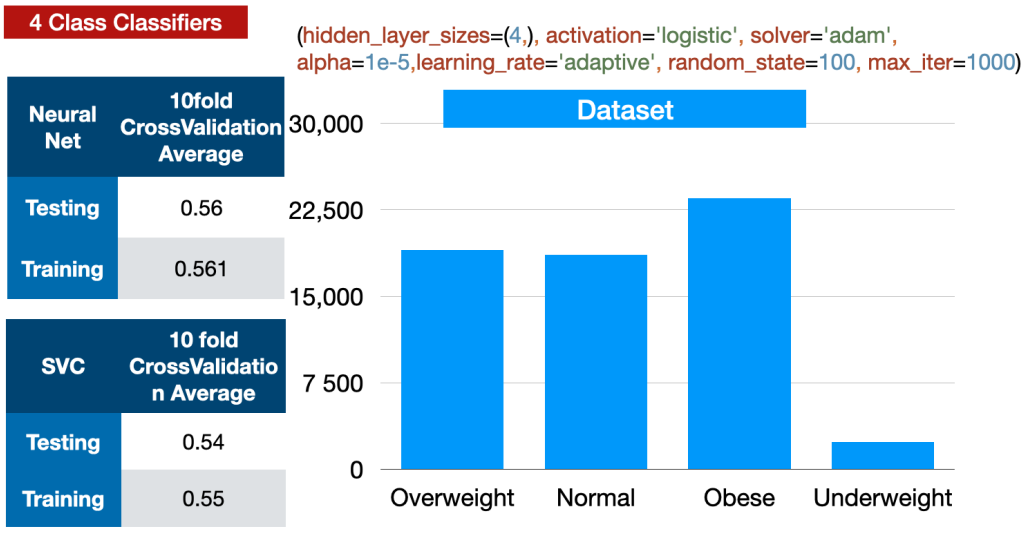

Figure 5. Four-class classifier.

In other related works, a relation between electrocardiogram (ECG) and MetS has been suggested via deploying neural network-based classifiers [22]. A link between MetS and a variety of demographic data and specific blood tests has also been established with the

deployment of neural networks, while other statistical methods were also compared [27]. The prediction of MetS, using artificial neural networks and clinical data, has also been explored, where BMI, age, HDL and LDL were identified as defining factors [28]. In other studies, an extreme learning machine approach was explored to identify the overweight class by using blood exams [12] using a sample of 500 data points of men and women. In a recent paper, a general diagnosis of MetS was pursued using clinical symptoms integrated with physio-chemical indexes in a smaller study of 586 cases where 450 participants had MetS and 136 others did not [29]. What is worth noting is also a recent paper with an extensive literature review related to MetS statistics and machine learning paradigms [11].

More precise comparisons of machine learning methodologies for predicting MetS and BMI can be seen in Tables 2 and 3, respectively. The main difference between our proposed methodologies and methodologies previously followed by other authors is that the latter, even though technically sound, lack a solid medical corroboration. The comparison is limited to papers that follow a similar pattern to ours, as per feature selection. Important differences between this work and previous related works include the size of samples and the fact that samples were in most cases imbalanced and could, thus, result in biased results, as outlined in Section 4. Another important factor that can add usability to a system is a feature selection methodology. We have ensured that our classification method explores methods that can assist medical prognosis in a novel manner by not using major defining factors that are already explored (e.g., triglycerides, cholesterol, BMI and glucose), as is the case in previous related works described in Tables 2 and 3.

Table 2. Comparison of methodologies— MetS.

Metrics	Cascaded SVC	QPHR1 [30]	ANN [28]
Accuracy (%)	84	>95	>95
Size of sample	5000	15,000	410
Balanced sample	yes	no	no
Features	Standard Biochemistry profile. Mets defining factors not included	Uses defining factors to predict MetS (triglycerides, Bmi)	Uses defining factors to predict MetS (BMI, diastolic blood pressure, HDL-cholesterol, LDL-cholesterol)

Table 3. Comparison of methodologies— BMI.

Metrics	4 Class Neural Network [5]	4 Class SVC	3 Class Neural Network [6]	3 Class SVC	Cascaded SVM	Extreme LM [12]
Accuracy (%)	56	55	58	62	85	90.54
Size of sample	75,000	15,000	33,000	15,000	10,000	500
Balanced sample	no	no	yes	yes	yes	no
All Classes?	yes	yes	no	no	yes	no (Overweight class)
Features	Full Biochemistry Profile	Full Biochemistry Profile	Full Biochemistry Profile	Full Biochemistry Profile	Without defining factors of MetS (15 features)	Blood Indexes (39 features) + age

Our aim is to create a streamlined machine learning pipeline that can cover as big a part of the patient's journey, minimize manual interventions and improve outcomes via minimizing required data to be fed in the process. We should also mention that for our experiments, a very broad dataset has being utilized of about 70,000 data points in both

balanced and imbalanced states. Clearly, this is a great increase over previous related studies that have incorporated small samples of an average of 500 imbalanced data points.

4. Exploratory Data Analysis and Bias Evaluation

Initially, the data were analyzed as per the BMI via deploying frequency histograms and probability density distributions to define how the sample used is represented. Data were also analyzed via deploying the same techniques as per each variable (blood exams—standard biochemistry profile) to determine average values, correlation between each variable and other statistical metrics, as in Figure 6. Variables were tested in both a generic way (complete data-set) and a more precise way as per weight category.

Key observations were drawn via examining important statistical metrics (e.g., mean, median, mode or standard deviation) to better understand the sample set under examination, to create the basis of available tools for future research endeavors and to develop a tool for missing value prediction through entity alignment. The results of this analysis can be seen in Figure 1. Clearly, the sample is well-balanced among the two genders and the age classes appear normally distributed. Thus, we can safely conclude that the network fed with this particular data is less likely to be biased, since gender is equally represented and a representative group is used for all age classes.

When examining the biochemistry profile, we observe mostly normally distributed values among all variables, as can be seen in Figure 7. The long tails observed in some variables suggest that more investigation could lead to the retrieval of valuable information.

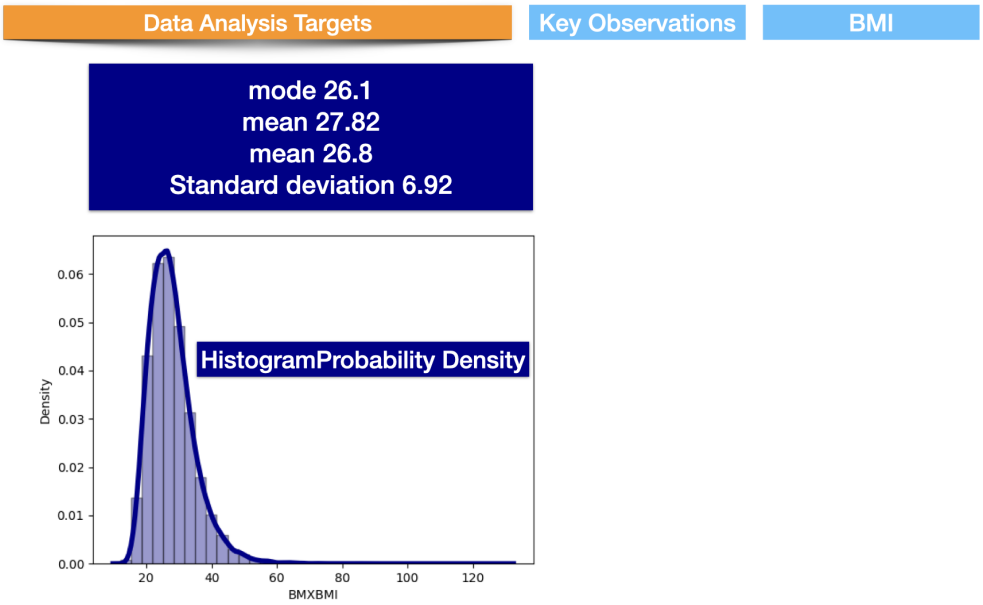

Figure 6. Key observations BMI.

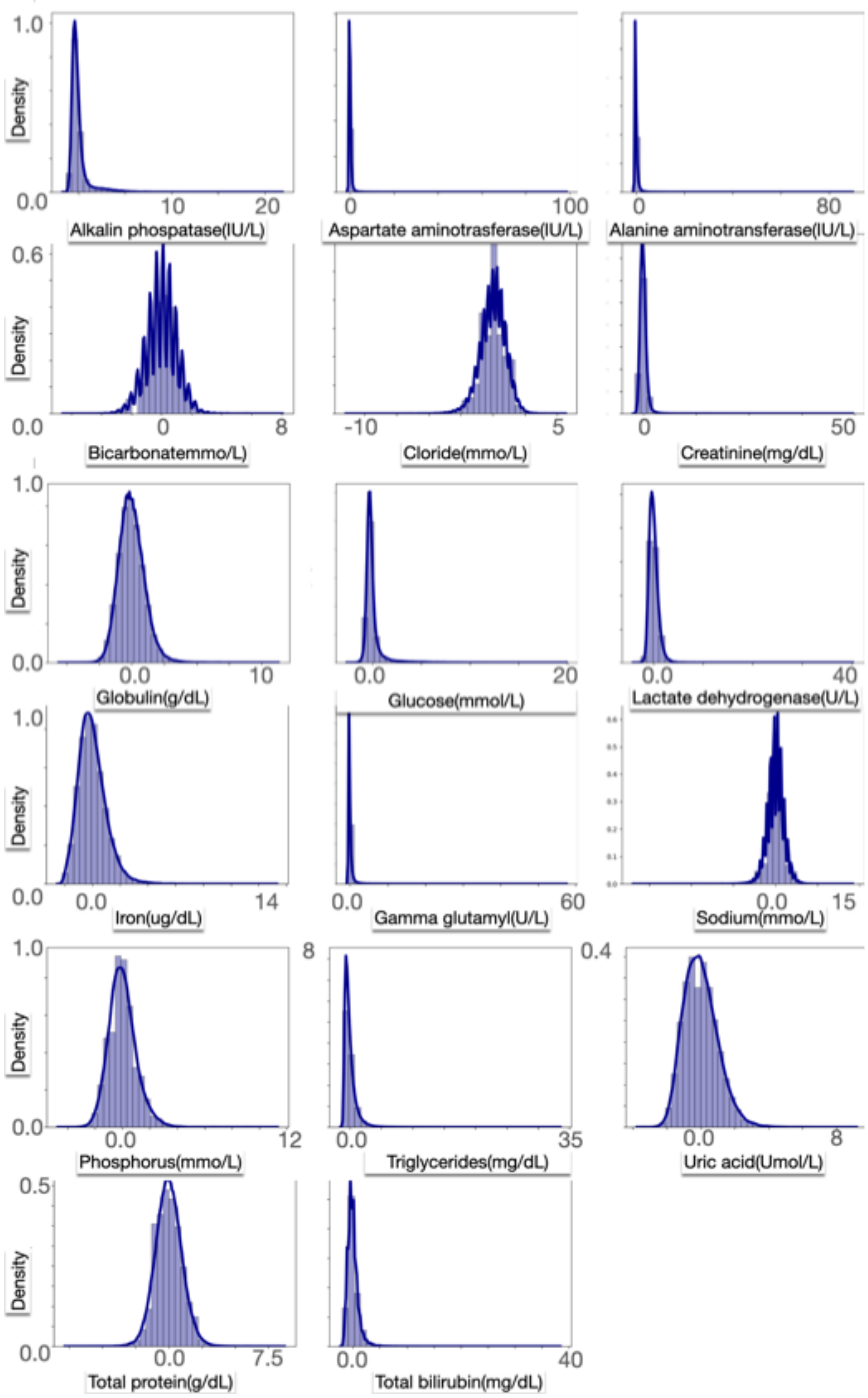

Figure 7. Histograms and densities of blood variables.

5. Predicting BMI Based on Blood Exams

5.1. Neural Networks vs. SVCs, Challenges and Objectives

We begin by comparing SVC and neural network-based classification accuracy in predicting BMI using blood exams. The tests were conducted via splitting the dataset in four and three weight classes. The accuracy of both systems is similar, and there is a slight increase when the testing is restricted to three classes via removing the underweight outlier for which fewer data-points are available. While on three classes, data were also balanced (Figure 8). The fact that a class was much less defined due to lack of data led us to deploy a one-class SVM to test the ability of the system to accurately define classes based on blood exams for each case studied.

The main challenges we faced in our previous studies was firstly the lack of data for some classes, which we addressed by balancing the data equally during examination. Another challenge was to find a way to limit classifications processes, in between classes and by limiting the cost of resources (computing power and time) in order to reach a conclusion and produce a classification result. This we achieved by using cascaded support vector classifiers, as discussed in Section 5.4.

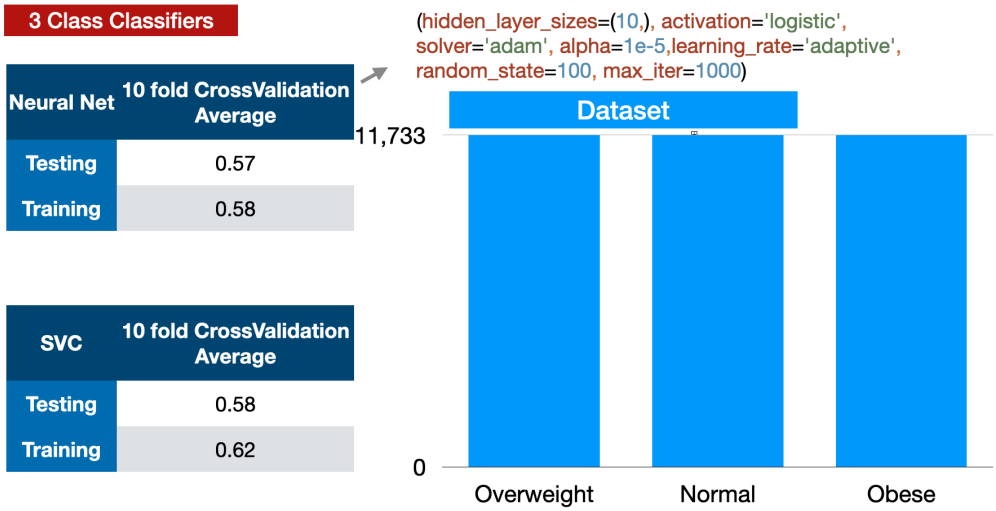

Figure 8. Three-class classifier.

5.2. One-Class SVM Identifiers

The dataset, the same as earlier on, was tested as per weight category. The corresponding results are summarized as the number of correct and incorrect predictions both in known (i.e., training) and unknown (i.e., testing) data. As seen in Figure 9, the classifier can easily correctly identify the class examined when data are unknown (testing sample). To evaluate identification accuracy, the tested sample included random data points from all classes to ensure high quality results and to simulate real world circumstances. The test and evaluation results presented in Figure 10 are an indicator of a well-defined sample of blood exams that can easily identify patterns and classify as per the task set for the system. In all cases, precision was estimated to about 92–95% with a nu parameter of 0.05.

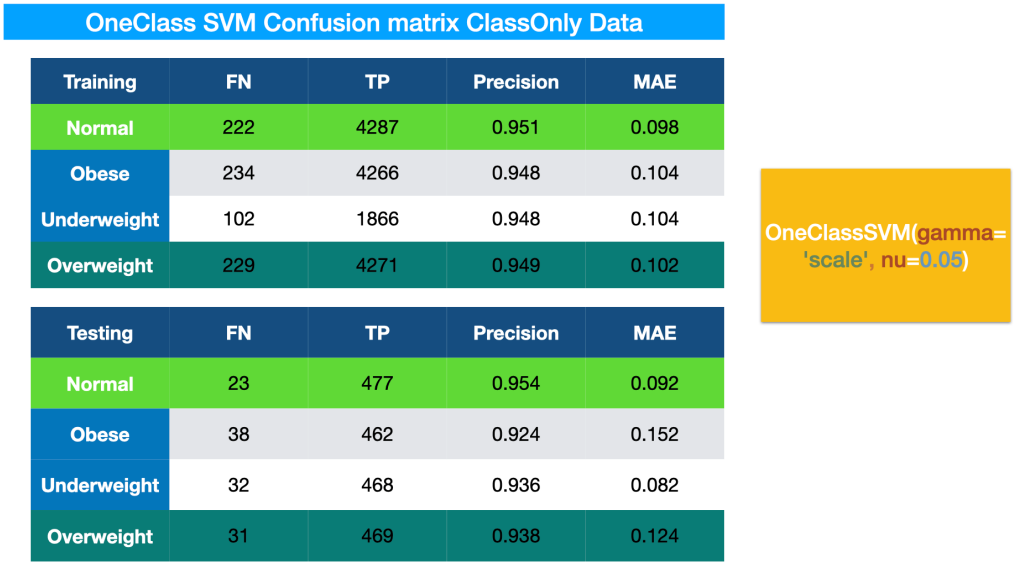

Figure 9. One-class SVM.

Figure 10. Confusion matrices.

5.3. One-vs.-One SVC

An ensemble machine learning model combines the predictions from multiple other models. It is a technique that may be used to improve model performance, ideally achieving better performance than any single model used in the ensemble.

A voting ensemble works by combining the predictions from multiple models. It can be used for classification or regression. In the case of regression, this involves calculating the average of the predictions from the models. In the case of classification, the predictions for each label are summed and the label with the majority vote is returned as the final and overall prediction [31].

As observed in Figure 11, there is a tremendous increase in classifying accuracy with a total average of 82% (for all cases calculated) when testing data are examined. The one-vs.-one SVC classes in Figure 12 can formulate the basis of a hard voting ensemble for categorising inputs as per weight category. However, a cascaded classifier was eventually preferred and selected as it is shown in the following Section 5.4 to perform significantly better.

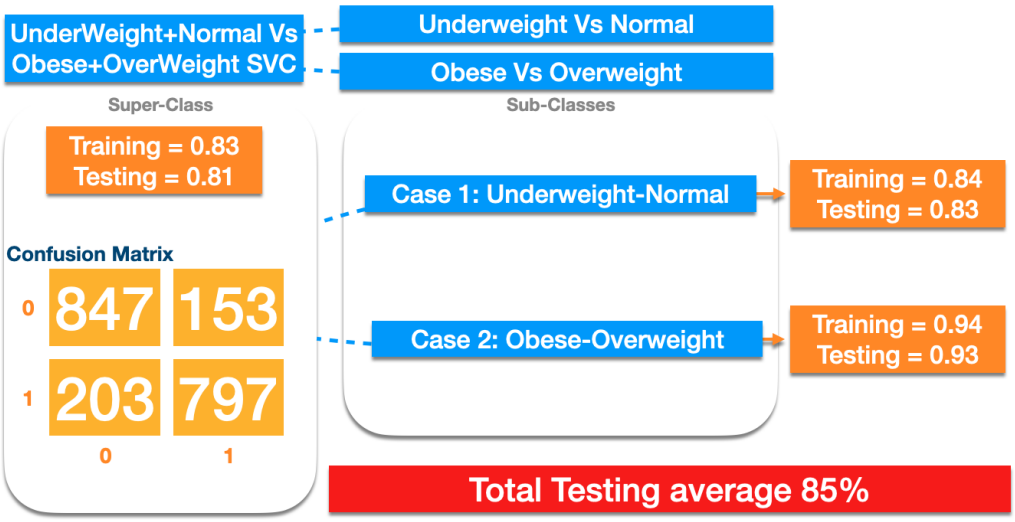

Figure 11. Cascaded SVC system.

OneVsOne NeuralNet classifier Accuracy 10 fold Cross Validation

Testing	OverWeight	Normal	UnderWeight	SubTotal Average
Obese	0.93	0.80	0.90	0.88
OverWeight		0.60	0.90	0.75
Normal			0.85	0.85
			TotalAverage	0.82

Training	OverWeight	Normal	UnderWeight	SubTotal Average
OBESE	0.95	0.79	0.95	0.89
OverWeight		0.64	0.90	0.77
Normal			0.86	0.86
			TotalAverage	0.84

Figure 12. One-vs.-one classifier.

5.4. Cascaded SVC

Since the one-vs.-one SVC was found to be promising, we developed and tested cascaded variations of it. Indeed, a one-vs.-one classifier cannot function as a standalone classifier and must be considered as a testing substructure for a system to build upon. Since a voting system could prove to be more costly and complicated, a cascaded methodology using SVCs was developed, as in Figure 11.

Firstly, two larger data groups (super-classes) were considered. The first group contains samples from both the underweight and normal classes, while the second group contains samples from both the overweight and obese classes. In this grouping, the samples in the extreme classes are combined with the samples of their corresponding neighboring classes, forming now only two major (super-)classes as opposed to the previous four weight classes. Next, when a binary classifier returns one of the two major classes, a second classifier returns a finer classification into one of the two (super-)classes that constitute it. That is, when the first classifier returns the combined underweight-normal class, the second classifier returns either the underweight or the normal class. Similarly, when the first classifier returns the combined overweight–obese class, the second classifier returns either the overweight or the obese class. In this cascade (two-level) classification scheme, one out of four classes is eventually returned, similarly to the original one-vs.-one four-class

classifier described earlier in Figure 12. However, the cascade classifier is shown to perform significantly better when tested and evaluated. Indeed, the total testing average of all processes achieves an accuracy of 85%, as shown in Figure 11. In all scenarios, about 4000 data-points were used distributed equally between classes (i.e., 2000 overweight vs. 2000 normal weight) from which a 10% sample was retained for verification(testing sample). Experiments were run on Python. All data were scaled within the 0 to 1 interval and a radial basis function kernel was used. Using grid search for the regularization parameter(C) and gamma(G), we concluded that accuracy was maximized when C is 1 and G is equal to 0.05 or 1/n_features, where features are 17. The search was conducted in two stages. In the first stage (line one of Figure 13), we incremented each run by one (step). When C was closer to 1 and gamma closer to 0, the accuracy increased; thus, to limit the search loop, we redesigned the grid search where C was in the range of 0.9 and 1 and gamma between 0.025 and 0.225, as can be seen in Figure 13. In Table 4, we show the kernels tested and their respected accuracy.

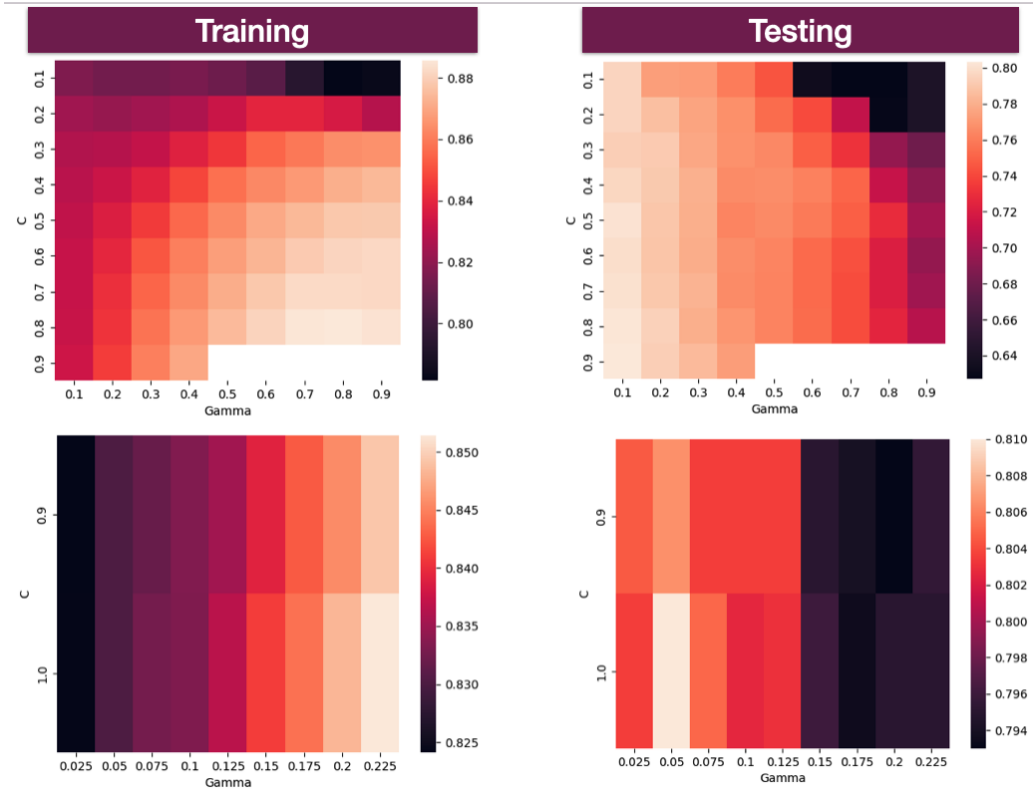

Figure 13. Grid search heatmap.

Table 4. Kernels in cascaded SVC and BMI super groups.

Set/Kernels	rbf	Linear	Poly	Sigmoid
training	0.83	0.804	0.8	0.7
testing	0.81	0.803	0.785	0.7

Even though the SVCs were the main approach used and thoroughly examined for the purposes of this research, other classifiers were also examined and tested. More precisely, Gaussian Naive Bayes, the Random Forest Classifier and Ada Boost were examined, as shown in Table 5. It is important to note that random forest classifiers showed promise and accuracy increased with increased test sample, but since they tend to over-fit we preferred SVCs as they better define separability between the classes. Cascaded SVCs have been proven to show great adaptability and accuracy in similar classification problems ([26]) and usability as a the base of recommender systems ([25]).

Table 5. Kernels in cascaded SVC and BMI super groups.

Set/Classifiers	Gaussian Naive Bayes	Random Forest Classifier	AdaBooster
training	0.7511	0.83	0.805
testing	0.732	0.72	0.781

6. Implementing a More Precise System

6.1. MetS Prediction Model

Via application of the same methodology as previously described, a model was deployed to classify the same population with regard to presence of MetS based on blood exams. As seen in Figure 14, triglyceride, cholesterol and glucose levels simultaneously being outside of normal/suggested ranges forms a defining factor of MetS. This is the, so-called, "rule of three". Ranges and criteria according to different methodologies and guidelines can be seen in Figure 15 ([32]).

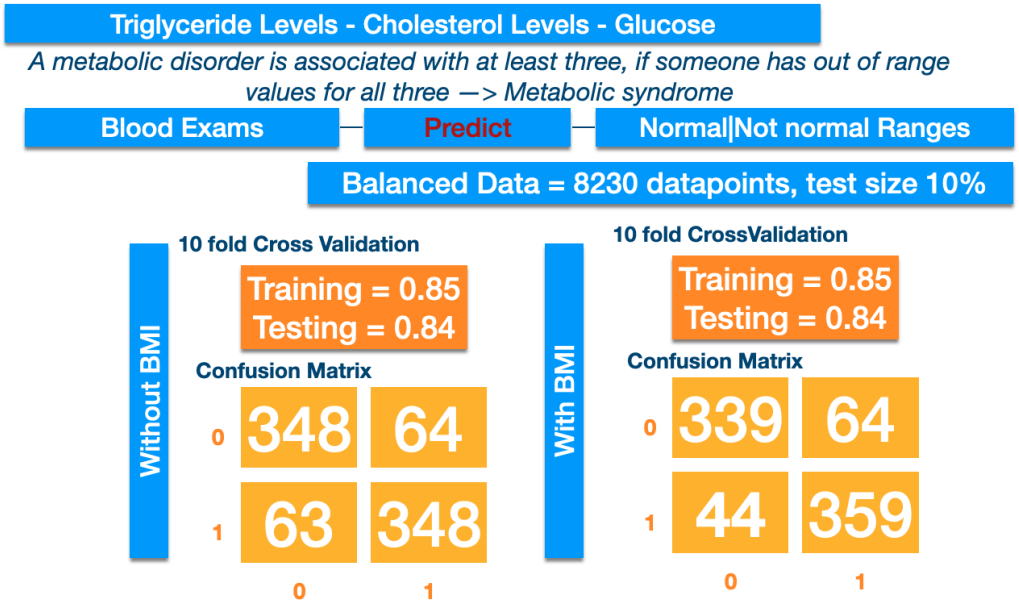

Figure 14. MetS classifier—total cholesterol.

	NCEP ATP III	WHO	EGIR	IDF
Required	None	Insulin Resistnace	Hyperinsulinemia	Central obesity
Criteria	**Any three of the five criteria**	**Insulin resistance or diabetes plus two of the five criteria below**	**Hyperinsulinemia plus two of the four criteria below**	**Obesity plus two of the four criteria below**
Obecity	Waist circumference	Waist to hip ratio	Waist circumference	Central obesity
Hyperglycemia	Fasting glucoce >=100 mg/dl or Rx	Insulin resistance	Insulin resistance	Fasting glucoce>=100mg/dl
Dyslipidemia	Triglycerides>=150mg/dl	Triglycerides>=150mg/dl	Triglycerides>177mg/dl	Triglycerides>150mg/dl
Dyslipidemia As a separate critereia	HDL cholesterol <40mgl/dl(M), <50mg/dl(F)	HDL cholesterol <35mgl/dl(M), <39 mg/dl(F)	HDL <39 mg/dl	HDL cholesterol <40mgl/dl(M), <50mg/dl(F)
Hypertension	>130 mmHg systolic|| 85>mmHg diastolic	>140/90 mmHg	>140/90 mmHg	>130 mmHg systolic|| 85>mmHg diastolic
Other		Microalbuminuria		

Figure 15. MetS guidelines and criteria.

To address this issue and implement a more precise and reliable system, we make the following four basic assumptions:

- While classifying for BMI, blood exams did not include triglycerides, glucose or cholesterol levels.
- Total cholesterol is computed using both LDL (bad cholesterol) and HDL (good cholesterol) and 20% of triglycerides. The out-of-bounds values of total cholesterol when, at the same time, triglycerides are out of range suggests that LDL is more prominent than HDL and that HDL has a greater probability of being low. To test this hypothesis, results need to be compared for total cholesterol or only HDL used as features.
- Since it was previously shown that BMI can be predicted quite accurately by using blood exams and BMI is a more common factor of MetS, there is good reason to believe that the classifier can identify patterns related to metabolism and, as such, be used in more refined calculations.
- The tests will not be gender specific. Any gender specificities will probably be identified by the model. Data will be balanced, but the road is always open for further experiments to be carried out in the future.

6.1.1. Total Cholesterol as a Feature

The SVC system has returned an average 10-fold classification accuracy in a tested sample of 84%. The hypothesis was also tested (Figure 16), and it verified that LDL is more prominent than HDL and that HDL has a greater probability of being low when total cholesterol and triglycerides are simultaneously out of bounds. This became evident as the prediction of BMI as a feature is not affected by the use of either the total cholesterol or only HDL, as seen in Figure 14.

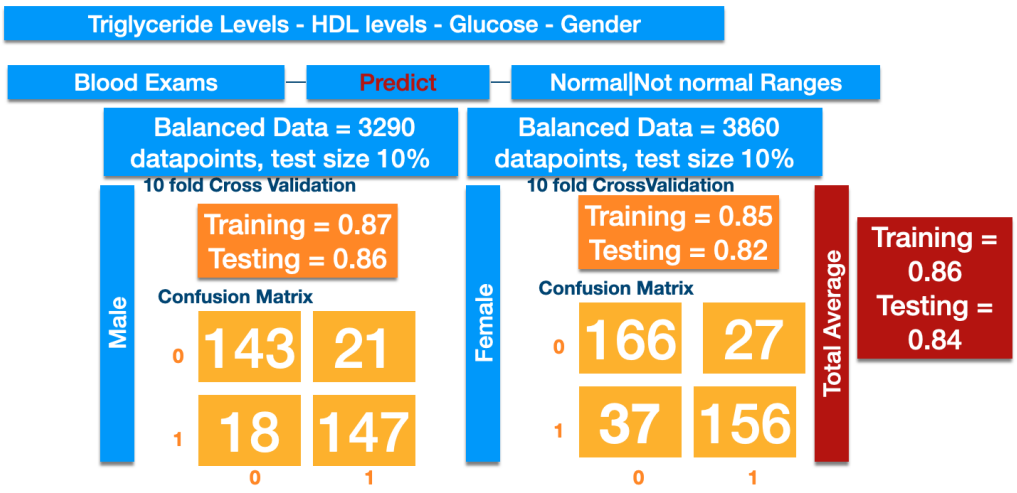

Figure 16. MetS classifier— HDL.

6.1.2. HDL as a Feature

In Figure 16, it is shown that the use of only HDL as a predictor instead of the total cholesterol returns almost identical results. This suggests that either variable could be utilized for classification purposes. Since the value ranges of HDL are dependent on gender [33], the hypothesis is tested separately for females and males and confirmed for both. Indeed, we performed a 10-fold cross validation run on a 10% testing sample of the about 3500 total samples used in both cases. The corresponding results of the classifier return an average classification accuracy of 86% for the male sample and 82% for the female sample and a total (over both males and females) average accuracy of 84%, as seen in Figure 16.

6.2. Blood Pressure Predictor

Finally, to complete our system, we enhanced it with a preliminary blood pressure predictor, which we plan to expand and improve further in a future system version. In corresponding experiments, we implemented optimized methods for increasing accuracy. The results can be seen in Figure 17. Specifically, systolic blood pressure (SBP) was predicted within and outside of normal ranges. Blood exams, triglycerides, total cholesterol and BMI were used as features. The 10-fold cross validation accuracy in the testing sample was 74%, which is very encouraging and will be improved further.

It is important to note that when using the standard biochemistry profile, we tested three different scenarios. For the first scenario, we used only blood exams, excluding the defining factors of MetS used in this study, i.e., triglyceride, glucose and HDL or total cholesterol measurements. In the second scenario, we used the full biochemistry profile, while in the third scenario, we also included BMI. The third scenario prevails by about 7–8% in the 10-fold cross validation cycle when all other parameters are equal.

Since the third scenario is much better in predicting SBP, it is the one used and the one proposed in the next section. By having MetS and BMI predicted per class as described in the previous steps, the proposed methodology creates a valuable tool that can be used when important values are missing.

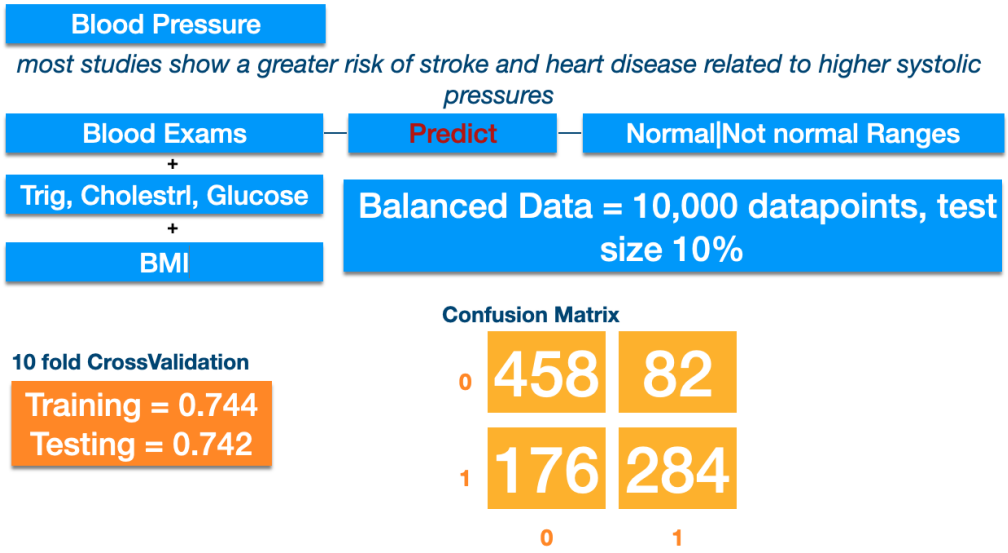

Figure 17. Blood pressure classifier.

7. Summary, Discussion, Itemization of Key Findings and Contribution

In this paper, we proposed a system (of superior performance) to predict MetS from blood exams as in Figure 18. Specifically, blood exams were used as input, but our efforts focused on the use of as small a number of parameters as possible for the initiation of the classification process. The classifier in the system predicts the BMI class ("underweight", "normal", "overweight" and "obese") and MetS state ("MetS present" and "MetS not present") without using related factors.

Depending on identified states and using blood exams, the system classifies blood pressure state or some other health state that is related to MetS. The system returns related risk factors and related recommendations (e.g., diet suggestions, lifestyle changes or medical interventions more commonly used). We deployed one-class SVMs to test the validity of our hypothesis and, thus, the ability of the classifier to identify body weight factors in blood exams. Using cascaded classifiers, an average accuracy of 85% was achieved, and BMI classification as per weight class, for all weight classes, based on standard biochemistry profile was optimized. By validating our initial hypothesis, other pathways were explored via applying similar methodologies.

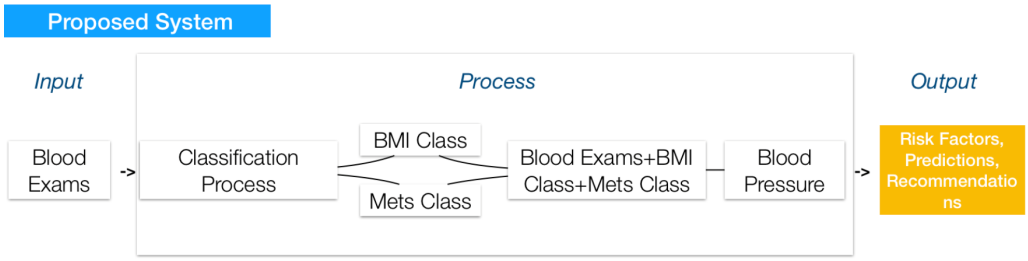

Figure 18. Combination of methodologies.

Weight being associated with metabolism and weight class being identified via blood exams, metabolic syndrome became the new classifying target. Using factors related to MetS, a new system was engineered that can identify this particular health state with an accuracy of 84%. Total cholesterol and HDL provided similar results when used alternately as factors of MetS in the classifier, even though the medical literature suggests that HDL is a key biomarker for MetS when combined with triglycerides and glucose.

High blood pressure, being both a factor and an outcome of MetS, was tested in a similar fashion. At this stage, SBP was evaluated using a full biochemistry profile and BMI. The final outcome was a 74% classification accuracy. Testing different scenarios, as described in detail in previous sections, we concluded that by using all parameters, as seen in Figure 18, a robust increase of about 8% in accuracy was achieved compared to using only parts of the biochemistry profile. Having already expanded on methodologies to identify the other parameters, this system as a whole can predict accurately systolic blood pressure even when some values are missing(missing value prediction) by using the available blood exams to classify BMI and MetS classes and, thus, enhancing the system with increased accuracy. More research is being conducted on this and will be published elsewhere in the near future. Finally, a basic system has been conceptualized (in Figure 18) and will be deployed as an interface with the optimized classifier.

8. Conclusions and Future Research Avenues

We see this study as another stepping stone for more applied methodologies in the area of artificial intelligence that could suggest novel methods for health care interventions and optimization of outcomes in a variety of ways. Via streamlined processes in health state prediction and general pathology, based on fewer variables and equally fewer exams, this research could result in a decrease in both time and economic costs. Accumulated risks via outcome prediction could be better evaluated. Basic diagnostic exams could be utilized to define more complicated outcomes. The employment of streamlined pattern recognition and machine learning technologies in population health metric analysis can lead to novel biomarker discovery. Implementation of policies based on system recommendations as per modeled outputs can be more easily achieved. Child obesity and age-specific analysis form both important research avenues and it is our aim to pursue them further in the future. A bigger spectrum of missing value identification by using pattern recognition and regression analysis is also a point of interest and a current endeavor of ours. Other research studies are currently underway, and their results will be presented elsewhere in the near future.

Author Contributions: Doctoral research of D.P.P. under the supervision of G.A.T. and D.N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partly supported by the University of Piraeus Research Center.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Publicly available data were used, where anonymity was ensured by the provider of the data.

Data Availability Statement: All data are available in the digital library of the center for disease control and prevention (CDC) https://www.cdc.gov (accessed on 1 February 2022) from which the National Health and Nutrition Examination Survey (NHANES 2001-2018) was utilized [34] for the purposes of this research.

Acknowledgments: Theoretical/medical support and technical/medical advice as per the validity of our hypothesis was provided by the doctors of Dermacen S.A https://www.dermatologikokentro.gr (accessed on 1 February 2022). For the implementation of the project and the analysis of the data the scikit-learn, Python libraries were used [35].

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SBP	Systolic blood pressure
METS	Metabolic Syndrome
BMI	Body mass index
ASCVD	Atherosclerotic cardiovascular disease
HDL	High-density lipoprotein
LDL	Low-density lipoprotein
NCD	Noncommunable disease
SVM	Support vector machine
SVC	Support vector classifier
Trig	Triglycerides
Cholestrl	Cholesterol

References

1. Thomas, D.; Ravi, K. National health and nutrition examination survey: Sample design. *Future Healthc. J.* **2019**, *6*, 94.
2. Strimbu, K.; Tavel, J.A. What are biomarkers? *Curr. Opin. HIV AIDS* **2010**, *5*, 463. [CrossRef] [PubMed]
3. Panagoulias, D.P.; Sotiropoulos, D.N.; Tsihrintzis, G.A. Towards Personalized Nutrition Applications with Nutritional Biomarkers and Machine Learning. In *Advances in Assistive Technologies: Selected Papers in Honour of Professor Nikolaos G. Bourbakis—Vol. 3*; Tsihrintzis, G.A., Virvou, M., Esposito, A., Jain, L.C., Eds.; Springer: Cham, Switzerland, 2022; Volume 28, pp. 73–122.
4. Panagoulias, D.P.; Sotiropoulos, D.N.; Tsihrintzis, G.A. Nutritional Biomarkers and Machine Learning for Personalized Nutrition Applications and Health Optimization. *Intell. Decis. Technol.* **2021**, *15*, 645–653. [CrossRef]
5. Panagoulias, D.P.; Sotiropoulos, D.N.; Tsihrintzis, G.A. Nutritional Biomarkers and Machine Learning for Personalized Nutrition Applications and Health Optimization. In Proceedings of the Twelfth IEEE International Conference on Information, Intelligence, Systems and Applications, Chania, Greece, 12–14 July 2021; pp. 731–733.
6. Panagoulias, D.P.; Sotiropoulos, D.N.; Tsihrintzis, G.A. Biomarker-based Deep Learning for Personalized Nutrition. In Proceedings of the 33rd IEEE International Conference on Tools with Artificial Intelligence (ICTAI), Virtually, 1–3 November 2021; pp. 73–122.
7. Massaro, A.; Ricci, G.; Selicato, S.; Raminelli, S.; Galiano, A. Decisional Support System with Artificial Intelligence oriented on Health Prediction using a Wearable Device and Big Data. In Proceedings of the 2020 IEEE International Workshop on Metrology for Industry 4.0 IoT, Roma, Italy, 3–5 June 2020; pp. 718–723.
8. Usman, A.; Won, L.J.; Syed, M.B.H.; Taqdir, A.; Ali, K.W.; Sungyoung, L. The Impact of Big Data in Healthcare Analytics. In Proceedings of the 2020 International Conference on Information Networking (ICOIN), Barcelona, Spain, 7–10 January 2020; pp. 61–63.
9. Ralf, H.; Michael, G. *Artificial Intelligence Applications in Human Pathology*; World Scientific Pub Co Inc.: Singapore, 2022; 216p.
10. WHO. Obesity and Overweight. 2021. Available online: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (accessed on 9 June 2021).
11. Kakudi, H.A.; Kiong, L.C.; Moy, F.M.; Kau, L.C.; Pasupa, K. Diagnosis of metabolic syndrome using machine learning, statistical and risk quantifiaction techniques: A systematic literature review. *Malays. J. Comput. Sci.* **2021**, *34*, 221–241.
12. Huiling, C.; Bo, Y.; Dayou, L.; Wenbin, L.; Yanlong, L.; Xiuhua, Z.; Lufeng, H. Using Blood Indexes to Predict Overweight Statuses: An Extreme Learning Machine-Based Approach. *PLoS ONE* **2015**, *10*, e0143003.
13. Khatiwada, S.; Sah, S.K.; Kc, R.; Baral, N.; Lamsal, M. Thyroid dysfunction in metabolic syndrome patients and its relationship with components of metabolic syndrome. *Clin. Diabetes Endocrinol.* **2016**, *2*, 3. [CrossRef]
14. Sanisoglu, S.Y.; Oktenli, C.; Hasimi, A.; Yokusoglu, M.; Ugurlu, M. Prevalence of metabolic syndrome-related disorders in a large adult population in Turkey. *Clin. Diabetes Endocrinol.* **2016**, *6*, 92. [CrossRef]

15. Denson, J.L.; Gillet, A.S.; Zu, Y.; Brown, M.; Pham, T.; Yoshida, Y.; Mauvais-Jarvis, F.; Douglas, I.S.; Moore, M.; Tea, K.; et al. Metabolic Syndrome and Acute Respiratory Distress Syndrome in Hospitalized Patients With COVID-19. *JAMA Netw. Open* **2021**, *4*, e2140568. [CrossRef]
16. Dagan, S.S.; Segev, S.; Novikov, I.; Dankner, R. Waist, circumference vs. body mass index in association with cardiorespiratory fitness in healthy men and women: A cross sectional analysis of 403 subjects. *Nutr. J.* **2013**, *12*, 12. [CrossRef]
17. Grundy, S.M. Metabolic syndrome: A multiplex cardiovascular risk factor. *J. Clin. Endocrinol. Metab.* **2007**, *92*, 399–404. [CrossRef]
18. Cui, R.; Iso, H.; Yamagishi, K.; Saito, I.; Kokubo, Y.; Inoue, M.; Tsugane, S.; JPHC Study Group. High serum total cholesterol levels is a risk factor of ischemic stroke for general Japanese population: The JPHC study. *Atherosclerosis* **2012**, *221*, 565–569. [CrossRef] [PubMed]
19. Nathan, D.M.; Davidson, M.B.; DeFronzo, R.A.; Heine, R.J.; Henry, R.R.; Pratley, R.; Zinman, B. Impaired fasting glucose and impaired glucose tolerance: Implications for care. *Diabetes Care* **2007**, *30*, 753–759. [CrossRef] [PubMed]
20. American Heart Association. *What Is High Blood Pressure?* South Carolina State Documents Depository; South Carolina State Library: Columbia, SC, USA, 2017; pp. 565–5698.
21. Nilsson, P.M.; Tuomilehto, J.; Ryden, L. The metabolic syndrome—What is it and how should it be managed? *Eur. J. Prev. Cardiol.* **2019**, *26*, 33–46. [CrossRef] [PubMed]
22. Lim, C.; Kim, J.Y.; Nam, Y. ECG Signal Analysis for Patient with Metabolic Syndrome based on 1D-Convolution Neural Network. In Proceedings of the 2020 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 16–18 December 2020; pp. 731–733.
23. Artham, S.M.; Lavie, C.J.; Milani, R.V.; Ventura, H.O. Obesity and hypertension, heart failure, and coronary heart disease—Risk factor, paradox, and recommendations for weight loss. *Ochsner J.* **2009**, *9*, 124–132.
24. Lemieux, I.; Lamarche, B.; Couillard, C.; Pascot, A.; Cantin, B.; Bergeron, J.; Dagenais, G.R.; Després, J.P. Total cholesterol/HDL cholesterol ratio vs. LDL cholesterol/HDL cholesterol ratio as indices of ischemic heart disease risk in men: The Quebec Cardiovascular Study. *Arch. Intern. Med.* **2001**, *161*, 2685–2692. [CrossRef]
25. Lampropoulos, A.S.; Sotiropoulos, D.N.; Tsihrintzis, G.A. Cascade hybrid recommendation as a combination of one-class classification and collaborative filtering. *Int. J. Artif. Intell. Tools* **2014**, *23*, 2685–2692. [CrossRef]
26. Sotiropoulos, D.N.; Tsihrintzis, G.A. *Machine Learning Paradigms—Artificial Immune Systems and Their Applications in Software Personalization*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 118, p. 330.
27. Morteza, S.; Yadollah, M.; Anoshirvan, K.; Farzad, H. Comparison of artificial neural network, logistic regression and discriminant analysis methods in prediction of metabolic syndrome. *Iran. J. Endocrinol. Metab.* **2010**, *11*, 645–653.
28. Hiroshi, H.; Tetsuro, T.; Shigenari, H.; Toshifumi, H.; Ikuo, S. Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. *Comput. Biol. Med.* **2011**, *41*, 1051–1056. [CrossRef]
29. Xia, S.-J.; Gao, B.-Z.; Wang, S.-H.; Guttery, D.S.; Li, C.-D.; Zhang, Y.-D. Metabolic syndrome, Machine learning, Diagnosis model, Symptoms, Traditional Chinese medicine, Physicochemical indexes. *Biomed. Pharmacother.* **2021**, *137*, 111–367. [CrossRef]
30. Worachartcheewan, A.; Nantasenamat, C.; Isarankura-Na-Ayudhya, C.; Prachayasittikul, V. Quantitative population-health relationship (QPHR) for assessing metabolic syndrome. *EXCLI J.* **2013**, *12*, 569.
31. Dietterich, T.G. Ensemble methods in machine learning. In *Archives of Internal Medicine*; Springer: Berlin/Heidelberg, Germany, 2000.
32. Huang, P.L. A comprehensive definition for metabolic syndrome. *Dis. Model. Mech.* **2009**, *2*, 231–237. [CrossRef] [PubMed]
33. Cooney, M.T.; Dudina, A.; De Bacquer, D.; Wilhelmsen, L.; Sans, S.; Menotti, A.; De Backer, G.; Jousilahti, P.; Keil, U.; Thomsen, T.; et al. HDL cholesterol protects against cardiovascular disease in both genders, at all ages and at all levels of risk. *Atherosclerosis* **2009**, *206*, 611–616. [CrossRef] [PubMed]
34. Johnson, C.L.; Dohrmann, S.M.; Burt, V.L.; Mohadjer, L.K. *National Health and Nutrition Examination Survey: Sample Design, 2011–2014*; Current Opinion in HIV and AIDS; US Department of Health and Human Services, Centers for Disease Control: Oxfordshire, UK, 2014.
35. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.