2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA) | 978-1-6654-9810-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/INISTA55318.2022.9894186

Efficient Data-driven Machine Learning Models for Hypertension Risk Prediction

Elias Dritsas, Sotiris Alexiou, Konstantinos Moustakas
Department of Electrical and Computer Engineering
University of Patras, Greece
Email:{dritsase, salexiou}@ceid.upatras.gr, moustakas@ece.upatras.gr
http://www.vvr.ece.upatras.gr/

***Abstract*—Hypertension is a chronic condition characterized by high pressure in the arteries of the human body. As a result, the heart is forced to work more intensively for the normal circulation of blood in the body. It is one of the most important risk factors for future fatal and non-cardiovascular diseases, stroke and kidney failure. In this article, Machine Learning (ML) is used to design effective models for predicting the long-term risk of older participants (over 50 years old) being diagnosed with hypertension. Our purpose is to train models with high sensitivity in identifying subjects at risk to avoid the future development and occurrence of hypertension following the proper interventions. In the context of the adopted methodology, two different class balancing methods are considered, under which features ranking is applied, and two ML models (namely, Decision tree and Naive Bayes) are compared based on Precision, Recall, F-Measure, Accuracy and Area Under Curve (AUC).**

***Index Terms*—Hypertension, blood pressure, prediction, machine learning**

I. Introduction

Blood pressure is the pressure exerted by the blood on the walls of the arteries and depends on the pulse volume (i.e. how much blood our heart expels in each contraction) and vascular resistance blood flow. Blood pressure is measured by two numerical indicators, one is the systolic pressure, and the other is the diastolic pressure. Systolic pressure indicates the pressure-tension exerted by the blood on the walls of blood vessels when it leaves the heart, while diastolic pressure expresses the pressure exerted by the blood on the walls of blood vessels when the heart dilates to refill with blood. The units of pressure are millimetres of mercury (mmHg) [1].

According to the World Health Organization, the normal blood pressure value of an adult must be less than 140/90 mmHg. Specifically, the systolic pressure should not exceed 140mmHg, and the diastolic should not exceed 90mmHg. These values are the limits for the definition of hypertension [2].

Hypertension is a disease of the heart and blood vessels and is without exaggeration a scourge of modern society. Its exacerbation is due, on the one hand, to the ageing of the population and, on the other hand, to the modern habits and tendencies of people [3].

More specifically, according to studies, many factors contribute to high blood pressure. A sedentary lifestyle and lack of physical exercise lead to obesity. In addition, eating unhealthy foods rich in salt and fat is a risk factor. Consumption of caffeine and alcoholic beverages increases the risk of hypertension. Finally, smoking and stress aggravate the condition [4]–[6].

In 95% of patients, hypertension is characterized as idiopathic, as it can not be attributed to a known pathological cause [7]. When there is a cause of hypertension (diseases of the kidneys, blood vessels, heart, thyroid, adrenal glands), then we refer to secondary hypertension [8].

Early diagnosis of hypertension is important to prevent heart attack or stroke as well as damage to organs such as the heart, brain, and kidneys [9]. In this direction, the science of medicine collaborates with data science. The techniques of artificial intelligence and machine learning have a significant contribution to the development of optimal prediction models [10] for various diseases, such as type 2 diabetes [11]–[14]), cholesterol [15], [16], sleep disorders [17], CVDs [18], COPD [19], stroke [20] and Covid-19 [21], etc. Besides previous diseases, many studies have been conducted for hypertension, which will be the issue of interest in this study.

In [22], the authors present a neural network model in order to predict hypertension and achieve this with an accuracy of 82%. Moreover, in [23], a series of Machine Learning prediction models concerning AUC, Sensitivity and Specificity are applied, and the stacking ensemble is selected as the best performer. Besides, the authors in [24] compare k-nearest neighbors (k-NN), support vector machine (SVM) with radial basis kernel function, linear and quadratic discriminant analysis (LDA), decision trees (DT), and naive Bayes (NB) classifiers for the arterial hypertension diagnosis. The LDA achieves the highest classification accuracy. Finally, in [25]. the authors used four classification algorithms (SVM, DT implemented by C4.5 algorithm, random forest (RF), and extreme gradient boosting) to predict if a participant has hypertension or not. The extreme gradient boosting has the best prediction performance with accuracy, F1, and AUC equal to 94.36%, 0.875, and 0.927, respectively.

The current work analyzes the risk factors of hypertension and presents the main steps of the adopted methodology. Specifically, data sampling techniques (random undersampling and oversampling based on Synthetic Minority Oversampling Technique (SMOTE)) are exploited in this study for balancing

class distribution. Decision Trees and Naive Bayes are utilized with different performance measures to evaluate their predictive ability. A public dataset has been exploited to validate the models' performance. In parallel, the same models will be assessed as part of the GATEKEEPER [26] project with pilot data.

The rest of the paper is structured as follows. In Section II, a brief description of the GATEKEEPER system is made. Moreover, in Section III, we describe the dataset and its features. The methodology we followed is reflected in section IV. Besides, in Section V, we discuss the obtained research outcomes. Finally, a summary of the results and future directions are mentioned in Section VI.

II. The GATEKEEPER System

The main objective of GATEKEEPER is to enable the development of a smart digital platform that connects healthcare providers, businesses and elderly citizens in order to promote healthier independent lives for the ageing population. For this purpose, advanced Information and Communications Technologies (ICTs) are combined and applied. An aspect of the system is the evaluation and incorporation of algorithms from the domain of AI and ML to be used as part of its interventions. The development of predictive models aims to early predict a personalised risk based on data from the pilots of the project. Hypertension is among the conditions that will be investigated in the GATEKEEPER.

III. Dataset Description

The present research work exploits a dataset derived from the Kaggle website. More specifically, the number of participants is 848. Each instance of the dataset is described by 13 attributes, which are fed into ML models, and 1 attribute that represents the target class. The attributes are analyzed as follows:

- **Age** (years): This variable captures the participant's age targeting those who are older than 50 years.
- **Gender**: This variable indicates the participant's gender. The percentage of males and females is 51.4% and 48.6%, respectively.
- **BMI** (Kg/m^2) [27]: This variable denotes the participant's body mass index.
- **Smoking** [5]: This variable captures the smoking habits of a participant (smoker, non-smoker). 52.7% of participants are smokers.
- **Daily steps**: This variable captures the number of average daily steps taken by the participant.
- **Daily alcohol** (ml): This variable captures the participant's average daily alcohol consumption.
- **Daily salt** (gr): This variable shows the participant's average daily salt consumption.
- **Stress Level** [4]: This variable shows the participant's stress level, which is captured into 3 categories (high 35.9%, medium 32.4% and low 31.7%).

TABLE I
Class distribution per sampling method

	No Sampling	Undersampling	Oversampling
Hyp	394	424	457
Non-Hyp	454	424	454
Total	848	848	911

- **CKD** [28]: This variable shows if the participant suffers from chronic kidney disease or not. The CKD prevalence in the dataset is 47.2%.
- **Hb** (mg/dl) [29]: This variable captures the level of hemoglobin (a protein in red blood cells).
- **Adrenal and thyroid disorders (ATD)** [30]: This variable shows if the participator suffers from adrenal and thyroid disorders. ATD's prevalence in the dataset is 43%.
- **SBP** (mmHg) [31]: It is the variable that captures the systolic blood pressure.
- **DBP** (mmHg) [31]: This variable captures the diastolic blood pressure.
- **Hypertension**: This variable shows whether a participator is hypertensive or not. In the following, the notation Hyp will refer to the hypertension class. 46.4% of participants have hypertension.

All variables are numeric apart from Gender, Smoking, Stress Level, CKD, ATD and Hypertension, which are nominal.

IV. Methodology

The adopted methodology consists of the following steps:

- Class Balancing
- Features Ranking
- Design of Classification Framework
- Models Evaluation

These stages will be analytically presented in the forthcoming sections.

A. Class Balancing

There are various methods to tackle the problem of non-uniform class distribution. In this study, we will focus on two well-known sampling approaches [32].

First, random undersampling is applied to target the majority class by randomly eliminating instances until to achieve the desired balance (or the instances are equal) in both classes. Also, in this study, SMOTE [33] was executed to increase the data of the minority class by 16%. It is an oversampling method that increases the data by creating synthetic data [34] on minority class using 5-NN classifier on the same features. It is used to trade-off between precision and recall or increase recall at the cost of precision. Class distribution is recorded in Table I, assuming no, under and oversampling cases. Also, the impact of sampling methods on the participants' distribution per age group and gender is depicted in Figures 1 and 2.

B. Features Importance

Features ranking will help us efficiently represent each record, focusing on those that will give more information about

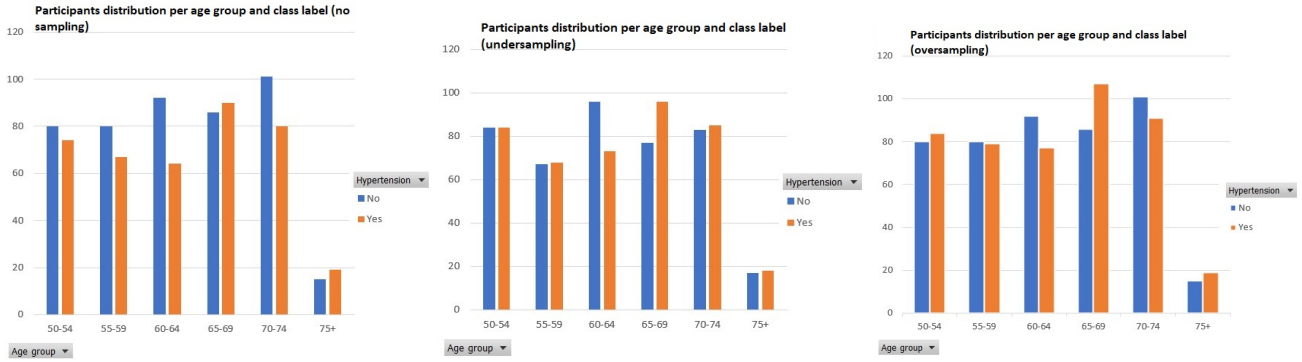

Fig. 1. Participants distribution per class and age group (no sampling, undersampling, oversampling).

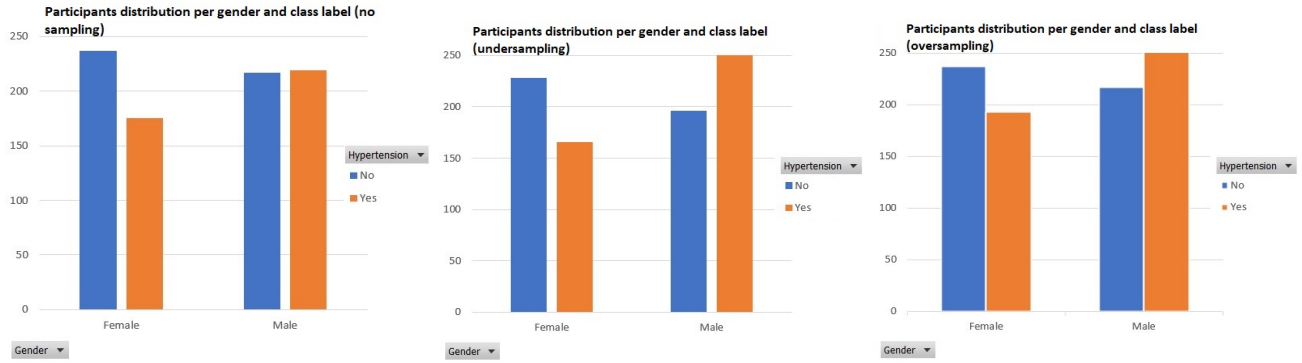

Fig. 2. Participants distribution per class and gender (no sampling, undersampling, oversampling).

the target class. For this purpose, Random Forest [35] was used to apply feature ranking. This method measures a feature's rank based on Gini impurity [36]. Table II shows the feature's importance in three different sampling cases.

In the original data, the ranker assesses as important features the SBP, DBP, CKD, ATD and Hb. Smoking and stress levels are ranked with zero importance. The rest features are of negative importance. These outcomes mean that these features don't contribute to the target class. However, salt and alcohol consumption are important risk factors for the occurrence of Hypertension. Actually, a diet with high salt and alcohol use impacts blood pressure and, thus, they are essential in the management of hypertension (treatment and control) [37]. The features' importance is differentiated between the sampling methods and the no sampling case. More specifically, after undersampling the initial dataset, daily salt consumption is fourth in order, and none of the features has a negative ranking. In oversampling case, the salt importance is positive but lower than the one in the undersampling case. Also, only stress and alcohol are of negative importance. At this point, we would like to note that all features will be considered to train and test the ML models.

C. Classification Framework

In healthcare, supervised learning has been extensively utilized to assess the risk for a disease manifestation using various features that capture the most important risk factors. Here, we deal with the design of ML models of high recall and

TABLE II
FEATURES IMPORTANCE BASED ON RANDOM FOREST AS RANKING METHOD

No Sampling		Undersampling		Oversampling	
Features	**Rank**	**Features**	**Rank**	**Features**	**Rank**
SBP	0.3148	SBP	0.3847	SBP	0.3532
DBP	0.2394	DBP	0.3106	DBP	0.2804
CKD	0.1804	Hb	0.2708	CKD	0.2206
ATD	0.1285	Daily salt	0.2528	ATD	0.1526
Hb	0.0771	Daily steps	0.2481	Hb	0.1115
Stress	0	CKD	0.1969	Gender	0.0483
Smoking	0	ATD	0.1580	Age	0.03535
Gender	-0.0035	Alcohol	0.1392	Smoking	0.03293
Age	-0.0071	Gender	0.0743	Daily steps	0.0206
Daily steps	-0.0230	BMI	0.0666	Daily salt	0.0143
Daily salt	-0.0307	Age	0.0271	BMI	0.0038
BMI	-0.0443	Stress	0.0212	Stress	-0.0018
Alcohol	-0.0856	Smoking	0.0165	Alcohol	-0.0555

AUC to ensure that new subjects can be correctly classified. For this purpose, we evaluate the prediction performance of Decisions Trees (DT) [38] and Naive Bayes classification methods.

A binary classification problem is formulated, i.e., the target class cl = "Hyp" (hypertension condition occurrence) or cl = "Non-Hyp" (non-occurrence of the hypertension condition). The features vector of a subject i is captured by $\mathbf{y}_i = \left[y_{i1}, y_{i2}, y_{i3}, \ldots, y_{in}\right]^T$, ensuring that the training data size is much higher than features dimension n.

The long-term risk prediction framework will consist of

TABLE III
PERFORMANCE METRICS DEFINITION

Metric	Hyp	Non-Hyp
Precision	$\frac{TP}{TP+FP}$	$\frac{TN}{TN+FN}$
Recall	$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$
F-Measure	$2\frac{Precision \cdot Recall}{Precision+Recall}$	
Accuracy	$\frac{TN+TP}{TN+TP+FN+FP}$	

J48, which is an extension of the ID3 DT [39]. A DT has a hierarchical tree structure, starting from a root node and, through decision rules, branches and intermediate nodes are derived until to reach the leaves of the tree. The tree-building process initiates by selecting the attribute with the highest gain ratio, the so-called splitting attribute. To overcome the existence of bias in features, its operation is based on a variation of information gains. A DT that is built upon a gain ratio has superior performance (accuracy) than using information gain in complex tasks. Following [39], the gain ratio (GR) of split S on feature j is defined as follows [40]:

$$GR(S,j) = \frac{Entropy\,(S)\sum_{s=1}^{l}(\mathrm{p_s} \times Entropy(\mathrm{p_s}))}{Spliting_{Info}} \quad (1)$$

Naive Bayes [41] is the second model that will be considered in the current study. It is a probabilistic classifier established on the Bayes theorem. The involved features in the model should be highly independent to ensure probability maximization. A new subject i is categorized to that class cl which ensures the maximization of the conditional probability $P\,(cl|\,y_{i1},\ldots,y_{in})$, defined as

$$\begin{aligned} P\,(cl|\,y_{i1},\ldots,y_{in}) &= \frac{P\,(y_{i1},\ldots,y_{in}|cl)\,P\,(cl)}{P\,(y_{i1},\ldots,y_{in})} \\ &= \frac{\prod_{j=1}^{n} P\,(y_{ij}|cl)}{P\,(y_{i1},\ldots,y_{in})}. \end{aligned} \quad (2)$$

In (2), $P\,(y_{ij}|cl)$ is the probability of the feature assuming class, and $P\,(y_{i1},\ldots,y_{in})$ and $P(cl)$ are the prior probabilities of the features and the class, correspondingly. The class label of an unknown instance is estimated by solving the following maximization problem

$$\hat{cl} = \arg\max P\,(cl)\prod_{j=1}^{n} P\,(y_{ij}|cl)\,, cl \in \{Hyp, Non-Hyp\}. \quad (3)$$

D. Evaluation Metrics

In this sub-section, we will evaluate the aforementioned ML models' performance based on accuracy, precision, recall, F-Measure, and AUC [20]. The definition of the desired metrics is shown in Table III, where TP, TN, FP, FN stand for the true positive, true negative, false positive and false-negative, respectively.

Precision will show how many of the instances that are classified as hypertensive (non-hypertensive) indeed stemmed from this class. Recall indicates how many of the hypertensive (non-hypertensive) instances are correctly estimated. Also, AUC shows the probability of a model correctly distinguishing the Hyp from Non-Hyp instances. Its values are between 0 and 1. F-Measure is the harmonic mean of precision and recall. Finally, the accuracy shows the total classification performance for both hypertensive and non-hypertensive instances.

TABLE IV
PRECISION OF J48 TREE

J48	Precision		
	No Sampling	**Undersampling**	**Oversampling**
Non-Hyp	0.890	0.926	0.900
Hyp	0.873	0.942	0.894
Average	0.882	0.934	0.897

TABLE V
RECALL OF J48 TREE

J48	Recall		
	No Sampling	**Undersampling**	**Oversampling**
Non-Hyp	0.890	0.943	0.892
Hyp	0.873	0.925	0.902
Average	0.882	0.934	0.897

V. RESULTS AND DISCUSSION

In this section, we will present the experiments' results which were derived in the WEKA [42] environment using 10-fold cross-validation. The models' efficiency will be assessed on the original dataset and its balanced versions. J48, implemented by the C4.5 algorithm, has been used to design the DT model. The J48 DT was configured as follows: confidenceFactor = 0.25, unpruned = false, minNumObj = 2 and binarySplitS = false.

In Tables IV-VII and VIII-XI, we summarize the metrics results of J48 and Naive Bayes, for both classes, under the three sampling cases. In the original dataset, the precision and recall of J48 have identical values in both classes. In the case of Naive Bayes, the precision of the Hyp class is 92.6%, which is 4.2% higher than the one of the Non-Hyp class. However, the recall of Non-Hyp is 94.1% which is higher by 8.3% than the Hyp class. Also, we observe that the undersampling method has improved the performance of both classification models in comparison with no sampling. Under this sampling method, the performance of J48 has benefited more than the Naive Bayes.

Moreover, in terms of oversampling, the recall and precision improvements are lower than the ones in the undersampling case. This behavior is observed in both models. At this point,

TABLE VI
F-MEASURE OF J48 TREE

J48	F-Measure		
	No Sampling	**Undersampling**	**Oversampling**
Non-Hyp	0.890	0.935	0.896
Hyp	0.873	0.933	0.898
Average	0.882	0.934	0.897

TABLE VII
AUC OF J48

J48	AUC		
	No Sampling	**Undersampling**	**Oversampling**
Non-Hyp	0.908	0.928	0.912
Hyp	0.908	0.928	0.912
Average	0.908	0.928	0.912

TABLE VIII
PRECISION OF NAIVE BAYES

Naive Bayes	Precision		
	No Sampling	**Undersampling**	**Oversampling**
Non-Hyp	0.884	0.893	0.880
Hyp	0.926	0.940	0.934
Average	0.904	0.916	0.907

it should be noted that J48 and Naive Bayes have the same recall of 90.2%, but the benefit in Naive Bayes is higher than J48 when compared to no sampling. Also, the recall of J48 in the Hyp class has been favored by the oversampled data without worsening the recall of Non-Hyp subjects and the precision of both. In Naive Bayes, there is a 4.4% increase in the recall of the Hyp class accompanied by a 0.8% increase in the precision of the same class. Besides precision and recall, a combinatory metric, F-Measure, has been recorded. This metric shows that both models are more efficient if they are trained with undersampled data.

Moving on to the ROC values, the AUC of J48 is the same in both classes. A similar trend is observed in Naive Bayes, which yields a higher AUC. Undersampling is the winner method. In either case, both models achieve values very close to 1. AUC reveals the models' capability to discriminate the hypertensive from non-hypertensive subjects. It is clear that, in the undersampling case, Naive Bayes achieves this with a high probability of 96.7%. Comparable performance is succeeded by J48, with an AUC of 92.8%.

The accuracy of J48 and Naive Bayes is illustrated in Figure 3. It is another metric that captures the total classification performance in both classes. Finally, taking into account the accuracy of the models with the above metrics, we conclude that in the specific data, the undersampling is more suitable for the design of the desired models.

TABLE IX
RECALL OF NAIVE BAYES

Naive Bayes	Recall		
	No Sampling	**Undersampling**	**Oversampling**
Non-Hyp	0.941	0.943	0.938
Hyp	0.858	0.887	0.902
	0.902	0.915	0.934

TABLE X
F-MEASURE OF NAIVE BAYES

Naive Bayes	F-Measure		
	No Sampling	**Undersampling**	**Oversampling**
Non-Hyp	0.911	0.917	0.908
Hyp	0.891	0.913	0.903
Average	0.902	0.915	0.906

TABLE XI
AUC OF NAIVE BAYES

Naive Bayes	AUC		
	No Sampling	**Undersampling**	**Oversampling**
Non-Hyp	0.965	0.967	0.963
Hyp	0.965	0.967	0.961
Average	0.965	0.967	0.962

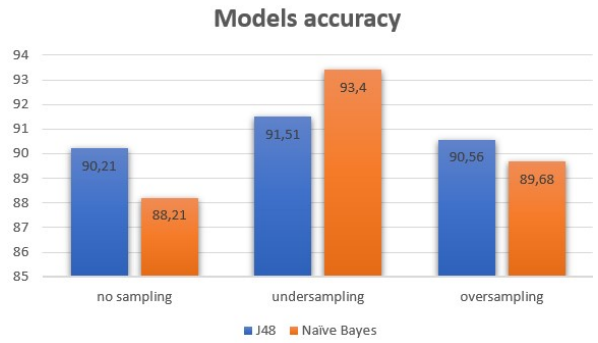

Fig. 3. J48 and Naive Bayes accuracy in terms of the sampling method

VI. CONCLUSIONS

In conclusion, a publicly available dataset was employed to analyse the importance of various risk factors for Hypertension. Then, these factors were considered to quantify the risk of hypertension occurrence, targeting people older than 50 years. We focus on this age group and especially those living at home to upgrade their quality of life through AI tools for personalized interventions. A framework with data-driven methods is suggested, and the role of class balancing (namely, random undersampling and oversampling) in feature ranking and ML methods performance was investigated. The results of both methods presented high recall and AUC, which prove that the models have high discrimination ability in identifying new subjects with Hypertension. Also, the accuracy and F-Measure reveal the overall classification efficiency of the models. In future work, we aim to experiment with more models as single classifiers (like SVM, Random Forest, Logistic Regression, Neural Networks) or apply ensemble learning techniques such as Stacking. Finally, our purpose is to emphasize class balancing and apply hybrid sampling methods before the evaluation of the ML models.

ACKNOWLEDGMENT

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.

REFERENCES

[1] M. Brunström and B. Carlberg, "Association of blood pressure lowering with mortality and cardiovascular disease across blood pressure levels: a systematic review and meta-analysis," *JAMA internal medicine*, vol. 178, no. 1, pp. 28–36, 2018.

[2] "World health organization: Hypertension," https://www.who.int/news-room/fact-sheets/detail/hypertension, (accessed on 1st June 2022).

[3] N. M. Kaplan, *Kaplan's clinical hypertension*. Lippincott Williams & Wilkins, 2010.

[4] S. Kulkarni, I. O'Farrell, M. Erasi, and M. Kochar, "Stress and hypertension." *WMJ: official publication of the State Medical Society of Wisconsin*, vol. 97, no. 11, pp. 34–38, 1998.

[5] A. Virdis, C. Giannarelli, M. Fritsch Neves, S. Taddei, and L. Ghiadoni, "Cigarette smoking and hypertension," *Current pharmaceutical design*, vol. 16, no. 23, pp. 2518–2525, 2010.

[6] E. S. Ford and R. S. Cooper, "Risk factors for hypertension in a national cohort study." *Hypertension*, vol. 18, no. 5, pp. 598–606, 1991.

[7] M. Wall, "Idiopathic intracranial hypertension," *Neurologic clinics*, vol. 28, no. 3, pp. 593–617, 2010.

[8] J. R. Chiong, W. S. Aronow, I. A. Khan, C. K. Nair, K. Vijayaraghavan, R. A. Dart, T. R. Behrenbeck, and S. A. Geraci, "Secondary hypertension: current diagnosis and treatment," *International journal of cardiology*, vol. 124, no. 1, pp. 6–21, 2008.

[9] S. Gulec, "Early diagnosis saves lives: focus on patients with hypertension," *Kidney international supplements*, vol. 3, no. 4, pp. 332–334, 2013.

[10] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, "An analytical method for diseases prediction using machine learning techniques," *Computers & Chemical Engineering*, vol. 106, pp. 212–223, 2017.

[11] N. Fazakis, O. Kocsis, E. Dritsas, S. Alexiou, N. Fakotakis, and K. Moustakas, "Machine learning tools for long-term type 2 diabetes risk prediction," *IEEE Access*, vol. 9, pp. 103 737–103 757, 2021.

[12] S. Alexiou, E. Dritsas, O. Kocsis, K. Moustakas, and N. Fakotakis, "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)*. IEEE, 2021, pp. 1–6.

[13] E. Dritsas, S. Alexiou, I. Konstantoulas, and K. Moustakas, "Short-term glucose prediction based on oral glucose tolerance test values," in *International Joint Conference on Biomedical Engineering Systems and Technologies - HEALTHINF*, vol. 5, 2022, pp. 249–255.

[14] E. Dritsas and M. Trigka, "Data-driven machine-learning methods for diabetes risk prediction," *Sensors*, vol. 22, no. 14, p. 5304, 2022.

[15] N. Fazakis, E. Dritsas, O. Kocsis, N. Fakotakis, and K. Moustakas, "Long-term cholesterol risk prediction with machine learning techniques in elsa database," in *13th International Joint Conference on Computational Intelligence (IJCCI)*. SCIPTRESS, 2021, pp. 445–450.

[16] E. Dritsas and M. Trigka, "Machine learning methods for hypercholesterolemia long-term risk prediction," *Sensors*, vol. 22, no. 14, p. 5365, 2022.

[17] I. Konstantoulas, O. Kocsis, E. Dritsas, N. Fakotakis, and K. Moustakas, "Sleep quality monitoring with human assisted corrections," in *International Joint Conference on Computational Intelligence (IJCCI)*. SCIPTRESS, 2021, pp. 435–444.

[18] E. Dritsas, S. Alexiou, and K. Moustakas, "Cardiovascular disease risk prediction with supervised machine learning techniques." in *ICT4AWE*, 2022, pp. 315–321.

[19] ——, "COPD severity prediction in elderly with ML techniques," in *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, 2022, pp. 185–189.

[20] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, p. 4670, 2022.

[21] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review," *Chaos, Solitons & Fractals*, vol. 139, p. 110059, 2020.

[22] D. LaFreniere, F. Zulkernine, D. Barber, and K. Martin, "Using machine learning to predict hypertension from a clinical dataset," in *2016 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2016, pp. 1–7.

[23] E. Dritsas, N. Fazakis, O. Kocsis, N. Fakotakis, and K. Moustakas, "Long-term hypertension risk prediction with ML techniques in ELSA database," in *International Conference on Learning and Intelligent Optimization*. Springer, 2021, pp. 113–120.

[24] V. S. Kublanov, A. Y. Dolganov, D. Belo, and H. Gamboa, "Comparison of machine learning methods for the arterial hypertension diagnostics," *Applied bionics and biomechanics*, vol. 2017, 2017.

[25] W. Chang, Y. Liu, Y. Xiao, X. Yuan, X. Xu, S. Zhang, and S. Zhou, "A machine-learning-based prediction method for hypertension outcomes based on medical data," *Diagnostics*, vol. 9, no. 4, p. 178, 2019.

[26] "Gatekeeper," https://www.gatekeeper-project.eu/, (accessed on 1st June 2022).

[27] M. Jiang, Y. Zou, Q. Xin, Y. Cai, Y. Wang, X. Qin, and D. Ma, "Dose–response relationship between body mass index and risks of all-cause mortality and disability among the elderly: a systematic review and meta-analysis," *Clinical Nutrition*, vol. 38, no. 4, pp. 1511–1523, 2019.

[28] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Ninchawee, "Predictive analytics for chronic kidney disease using machine learning techniques," in *2016 management and innovation technology international conference (MITicon)*. IEEE, 2016, pp. MIT–80.

[29] D. T. Gilbertson, J. P. Ebben, R. N. Foley, E. D. Weinhandl, B. D. Bradbury, and A. J. Collins, "Hemoglobin level variability: associations with mortality," *Clinical Journal of the American Society of Nephrology*, vol. 3, no. 1, pp. 133–138, 2008.

[30] T. de Silva, G. Cosentino, S. Ganji, A. Riera-Gonzalez, and D. S. Hsia, "Endocrine causes of hypertension," *Current Hypertension Reports*, vol. 22, no. 11, pp. 1–13, 2020.

[31] A. C. Flint, C. Conell, X. Ren, N. M. Banki, S. L. Chan, V. A. Rao, R. B. Melles, and D. L. Bhatt, "Effect of systolic and diastolic blood pressure on cardiovascular outcomes," *New England Journal of Medicine*, vol. 381, no. 3, pp. 243–251, 2019.

[32] T. Hasanin, T. M. Khoshgoftaar, J. Leevy, and N. Seliya, "Investigating random undersampling and feature selection on bioinformatics big data," in *2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2019, pp. 346–356.

[33] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using smote and effective data mining techniques," *IEEE access*, vol. 9, pp. 39 707–39 716, 2021.

[34] E. Dritsas, N. Fazakis, O. Kocsis, K. Moustakas, and N. Fakotakis, "Optimal team pairing of elder office employees with machine learning on synthetic data," in *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*. IEEE, 2021, pp. 1–4.

[35] Y. Liu, Y. Wang, and J. Zhang, "New machine learning algorithm: Random forest," in *International Conference on Information Computing and Applications*. Springer, 2012, pp. 246–252.

[36] C. Aldrich, "Process variable importance analysis by use of random forests in a shapley regression framework," *Minerals*, vol. 10, no. 5, p. 420, 2020.

[37] R. Gupta and S. Guptha, "Strategies for initial management of hypertension," *The Indian journal of medical research*, vol. 132, no. 5, p. 531, 2010.

[38] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[39] I. D. Mienye, Y. Sun, and Z. Wang, "Prediction performance of improved decision tree-based algorithms: a review," *Procedia Manufacturing*, vol. 35, pp. 698–703, 2019.

[40] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques," *IEEE Access*, vol. 7, pp. 1365–1375, 2018.

[41] S. Chen, G. I. Webb, L. Liu, and X. Ma, "A novel selective naïve bayes algorithm," *Knowledge-Based Systems*, vol. 192, p. 105361, 2020.

[42] "Weka tool," https://www.weka.io/, (accessed on 1st June 2022).