Automated Detection of Polycystic Ovary Syndrome Using Machine Learning Techniques

Yasmine A. Abu Adla
Electrical and Computer Engineering Department
Rafik Hariri University
Damour, Lebanon
abouadlaya@students.rhu.edu.lb

Dalia G. Raydan
Electrical and Computer Engineering Department
Rafik Hariri University
Damour, Lebanon
raydandg@students.rhu.edu.lb

Mohammad-Zafer J. Charaf
Electrical and Computer Engineering Department
Rafik Hariri University
Damour, Lebanon
charafmj@students.rhu.edu.lb

Roua A. Saad
Electrical and Computer Engineering Department
Rafik Hariri University
Damour, Lebanon
saadra@students.rhu.edu.lb

Jad Nasreddine
Computer and Information Systems Department
Rafik Hariri University
Damour, Lebanon
nasreddinejn@rhu.edu.lb

Mohammad O. Diab
Electrical and Computer Engineering Department
Rafik Hariri University
Damour, Lebanon
diabmo@rhu.edu.lb

2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME) | 978-1-6654-3501-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICABME53305.2021.9604905

***Abstract*— Polycystic Ovary Syndrome (PCOS) is a medical condition affecting the female's reproductive system causing ano/oligoovulation, hyperandrogenism, and/or polycystic ovaries. Due to the complexities in diagnosing this disorder, it was of upmost importance to find a solution to assist physicians with this process. Therefore, in this study, we investigated the possibility of building a model that aims to automate the diagnosis of PCOS using Machine Learning (ML) algorithms and techniques. In this context, a dataset that consisted of 39 features ranging from metabolic, imaging, to hormonal and biochemical parameters for 541 subjects was used. First, we applied pre-processing on the data. Hereafter, a hybrid feature selection approach was implemented to reduce the number of features using filters and wrappers. Different classification algorithms were then trained and evaluated. Based on a thorough analysis, the Support Vector Machine with a Linear kernel (Linear SVM) was chosen, as it performed best among the others in terms of precision (93.665%) as well as high accuracy (91.6%) and recall (80.6%).**

Keywords—PCOS, Machine Learning, Disease Detection, Classification

I. INTRODUCTION

Polycystic ovary syndrome (PCOS) is a complex heterogeneous endocrine disorder that affects the female's reproductive system [1]. PCOS usually affects 12 to 21% of reproductive-age women, out of which 70% remain undiagnosed [2]. This syndrome often causes menstrual abnormalities, hyperandrogenism, fertility problems, and metabolic sequelae. It is characterized by features that include irregular menstrual cycles, acne, and hirsutism [3]. Additionally, PCOS typically involves hormonal imbalances, insulin resistance, and metabolic abnormalities, which significantly increase the risk of infertility, type 2 diabetes, and cardiovascular disease [4].

Diagnosing PCOS can be a challenging process given the ambiguity of the symptoms. According to the Endocrine Society, clinicians are advised to diagnose PCOS using the Rotterdam criteria, which requires the presence of at least two of the following three medical conditions: hyperandrogenism, ovulatory dysfunction, and polycystic ovaries. Alongside these symptoms, a measure of the LH and FSH levels can be beneficial in establishing an LH/FSH ratio. In fact, a ratio greater than 2 usually indicates PCOS [5].

Despite the presence of diagnostic criteria, there are various clinical uncertainties and complexities when it comes to the diagnosis of PCOS. To begin with, very few research has been done on the diagnosis of PCOS given the ambiguity of its symptoms and their broad spectrum of severity. To exemplify, it has been reported that PCOS has multiple heterogeneous presentations that vary according to age and race, and can be indicative of multiple other comorbidities. Although physicians reported that they are familiar with the most common comorbidities, such as obesity, diabetes, insulin resistance, and depression, they are less likely to associate between PCOS and conditions such as sleep apnea, fatty liver, endometrial cancer, gestational diabetes, and anxiety. Therefore, due to its diverse presentation, diagnosing PCOS may need the evaluation of different physicians. However, the lack of coordination between doctors might result in the lack of a comprehensive assessment, which is needed to diagnose PCOS [6].

Furthermore, PCOS usually goes underdiagnosed by many healthcare providers. According to a community-based study, 18% of women had PCOS out of which 70% remain underdiagnosed. Moreover, current literature suggests that the under diagnosis of PCOS is exacerbated, especially in adolescents. Primary indicators of PCOS, such as anovulation, cyst development, presence of multiple follicles, hormonal fluctuations, and acne, are present during puberty as well.

Hence, physicians tend to be less concerned about these symptoms at that age [6]. Other concerns were raised regarding the over diagnosis of PCOS as well; as some clinicians expand the diagnostic criteria of PCOS to include polycystic ovaries [6].

Due to all above complexities, inconsistent approaches have been taken to diagnose women with PCOS. In fact, it has been reported that women were dissatisfied, frustrated, and confused with their diagnostic experience and had to wait an average of 2 years to receive the correct diagnosis [7]. A study conducted on patient satisfaction with PCOS diagnosis revealed that only 35.2% were satisfied with their experience, based on the time it took to diagnose and number of physicians seen before attaining a final diagnosis [6]. In fact, as many patients have to dedicate a lot of time to undergo several tests and visit several specialists to be taken seriously and get their symptoms checked out, many have resorted to self-diagnosis and managing PCOS on their own [6].

Thus, this research aims to build a high performing diagnostic model using Machine Learning (ML) in order to reduce and prevent human error that may affect PCOS diagnosis. By integrating ML into PCOS diagnosis, we can ease the lives of women suffering with PCOS, as a timely diagnosis can help in the early treatment of symptoms, infertility issues, and engagement in an adequate lifestyle to prevent weight gain and metabolic issues.

The remaining of this paper is organized as follows. Section II covers different ML techniques found in the literature to diagnose PCOS. Then, Section III tackles the proposed methodology. Section IV discusses the results followed by the conclusion and future perspectives.

II. Literature Review

A. Diagnosing PCOS using different ML techniques

A study was conducted in [8] on 200 women, where 150 were diagnosed with PCOS and the remaining 50 were not. Nine clinical and metabolic parameters (Age, BMI, LH, FSH, Systolic Blood Pressure, Diastolic Blood Pressure, Cycle Length, Fasting Blood Sugar, and Post Prandial Blood Sugar) were collected from the patients. To select the most significant features, a statistical analysis was done using two sample t-tests; the four features that had a p-value < 0.001 were considered significant. In order to classify patients into normal and PCOS groups, 2 classifiers were considered: Logistic Regression and Bayesian. The Bayesian classifier was able to Logistic regression (91.04%) when using 3-fold cross-validation.

A novel ML technique to diagnose PCOS using 18 parameters based on lifestyle and food intake was designed in [9]. All features were fed into several classifiers and the one with the highest accuracy was chosen. In fact, the Naïve Bayes algorithm gave the highest accuracy (97.65%) when tested on 119 samples using the hold out method.

Other researchers were able to achieve an accuracy of 89% using the Random Forest classifier [2]. The classifier was run on 541 subjects (364 normal and 177 PCOS) using a feature vector with 39 parameters. The number of features was reduced using Principal Component Analysis (PCA) test and inter-correlated features were merged. The remaining 23 features were then verified for their discriminating potential using SPSS. The final feature vector was run using the testing set (20% of the data) on Linear Regression (LR), K- Nearest Neighbors (KNN), Random Forest Classifier (RFC), and Support Vector Machines (SVM), etc. The RFC had the highest accuracy (89%) and was thus chosen.

Generally, most of the work done in this field went for the same processing and ML techniques to build a diagnostic model. Most researchers did not use all possible medical features and tested their results on a low number of subjects. However, in this study, we propose a PCOS diagnostic model based on medically reliable features. Additionally, our model was able to achieve high results while being tested on a high number of subjects. Finally, we considered precision as our decisive outcome based on physician recommendation.

This section tackles the proposed methodology for building a PCOS diagnostic model. Section A describes the dataset used. Then, the preprocessing steps are tackled in section B followed by feature selection in section C. Section D discuses training and evaluating the model.

III. Methodology

A. Dataset Description

The dataset at hand, which is entitled "Polycystic Ovary *Syndrome*", was found available in the datasets of the online community "Kaggle" [10]. The dataset contains clinical and physical parameters of 541 subjects out which 177 are iagnosed with PCOS and 364 are healthy. The dataset consists of 39 parameters ranging from metabolic parameters (9), physical parameters (16), and ultrasound-imaging parameters (3), to hormonal parameters (9) and biochemical ones (2). Some of these parameters include the patient's weight, height, age, blood group, blood pressure, cycle length, endometrium thickness, Follicle Stimulating Hormone (FSH) levels, etc. The data were collected from 10 different hospitals across Kerala, India.

B. Pre-processing

Electronic Health Records (EHR) data are usually "messy" in the sense that they can be incomplete, noisy, and inconsistent. Therefore, there is a need for meticulous data cleaning steps prior to any analysis [11]. Hence, the first step in any ML application is to pre-process the dataset at hand. Pre-processing involves data cleaning from any errors or missing values and performing different transformations and manipulations.

1) Detecting and Classifying Outliers and Errors

In medical records, outliers, the extreme values in a dataset are not always considered as errors, but rather an indication on the patient's condition; in such a case, the outlier may provide useful information regarding the patients' conditions. On the other hand, other values may point to human error, and hence the observation should be considered for removal or correction in order to avoid having a biased model. Therefore, we visualized the record of every patient based on the parameters in the form of a boxplot as shown in Figure 1. For each feature, we considered extreme values located outside the lower and upper fence of the boxplot and compared them to the parameter's normal range and the patient's condition. After consulting an obstetrician-gynecologist (OB-GYN), we classified the value as either an outlier or error. The outliers were kept in the dataset however; the subject that had an error in her information was dropped.

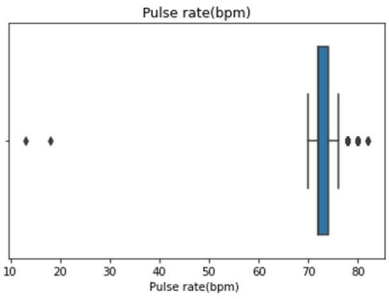

Figure 1. Boxplot for the Pulse Rate Feature

2) Standardization of the Data

After detecting the errors in the dataset and removing them, standardization of the dataset took place. Standardization of data is an essential step in pre-processing since it transforms the features to fit a common range so that

the larger numeric features do not dominate those with smaller values [12].

In this work, we used Z-score normalization as it keeps the statistical characteristics of the data. Feature records are transformed using Equation (1) to have a mean equals to zero and variance equals to one [12]. The equation of Z-score normalization is given below, where feature i records ($x'_{i,n}$) are transformed into records $x'_{i,n}$.

$$x'_{i,n} = \frac{x_{i,n} - \mu_i}{\sigma_i} \quad (1)$$

Where μ_i and σ_i are the mean and standard deviation of feature i, respectively.

C. Feature Selection

Feature selection (FS) is a crucial step to improve classifier's performance and reduce computation time. FS consists of selecting the most effective or prominent features from the entire set of features. When irrelevant or redundant features are removed, the performance of the classifier will be boosted and good classification results will be achieved [13].

In this study, we utilize a hybrid feature selection approach to select the optimal feature subset. A typical hybrid approach first applies filters to remove the least significant features and then applies wrappers to fine tune and further reduce the dimensionality of the dataset. Filters are models that depend on the training dataset rather than depending on the classifier or the predictor. Since filters do not depend on the classification algorithm, they are considered simple, fast, and can easily scale to high dimensional datasets [13]. On the other hand, wrappers select a subset of features based on the performance of the classifier. Unlike the filter approach, wrappers consider the correlation between features and the way they interact with the Machine Learning algorithm [13]. In our case, filters are applied first to reduce the number of features, which makes it less computationally expensive when wrappers are applied [14].

1) Filters

To perform feature selection on the continuous variables in our dataset, we applied the Analysis of Variance (ANOVA) test. ANOVA is a statistical test that compares the means of several samples. The ANOVA test reveals which features have great influence on the outcome or on other parameters [15].

After performing the ANOVA test on each feature with respect to our output (whether the patient suffers from PCOS or not), the p-value was computed. If the p-value was less than 0.05, the feature was considered significant and vice versa. The results obtained are shown in the Table I.

TABLE I. STATISTICALLY SIGNIFICANT FEATURES BASED ON THE ANOVA TEST

Feature	P-value
BMI	0.000012
Cycle Length	0.000011
Marriage Status	0.003796
FSH	0.005678
LH	0.01959
LH/FSH	0.00046
AMH	0.000000001416478
Left Follicle No.	1.900231E-51
Right Avg. Follicle Size	0.05398
Endometrium Thickness	0.02598

After performing the ANOVA test on each feature, we consulted an OB-GYN to make sure that the results obtained were aligned with medical field standards. The physician recommended that we keep some features that the ANOVA test labeled as insignificant since they are crucial for the diagnosis of PCOS (Table II). Therefore, in order to find a balance between the features recommended by the physician and those found significant by the ANOVA test, we decided to filter out the features that were considered insignificant by both the doctor and the statistical test, which are blood group, RR, FSH/LH, Vitamin D3, systolic blood pressure, and diastolic blood pressure.

TABLE II. FEATURE SIGNIFICANCE BASED ON ANOVA TEST AND DOCTOR RECOMMENDATION

Feature	Statically Significant	Doctor Recommendation
Cycle Length	Yes	Yes
Marriage Status	Yes	Yes (for delayed fertility)
No of abortions	No	Yes (PCOS causes miscarriages)
FSH	Yes	Yes
LH	Yes	Yes
LH/FSH	Yes	Yes
FSH/LH	No	No
TSH	No	Yes (for differential diagnosis)
AMH	Yes	Yes
PRL	No	Yes (for differential diagnosis)
Vitamin D3	No	No
PRG	No	Yes (gives an idea on ovulation)
RBS	No	Yes (gives an idea on the metabolism)

2) Sequential Forward Floating Feature Selection Wrapper

The Sequential Forward Floating Selection (SFFS) is a wrapper that starts with an empty set of features, and sequentially adds features that improve the performance of the classifier. The SFFS adjusts the trade-off between forward and backward steps, dynamically, by applying after each forward step a number of backward steps as shown in Figure 2. This is considered as 'self-controlled backtracking', where the resulting feature subsets result in a better classifier performance compared to the previously evaluated ones at that level. Consequently, there are no backward steps at all, if the result at that specific level does not yield better classifier performance [16].

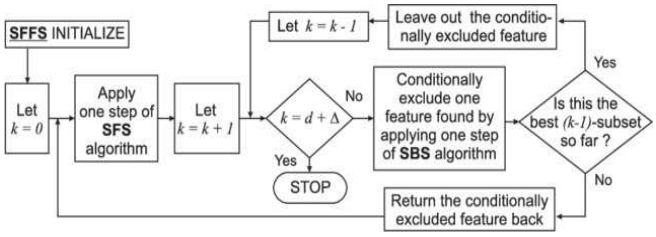

Figure 2. Sequential Forward Floating Selection Algorithm

The SFFS wrapper was applied on the following classifiers:

1) Logistic Regression (LR)
2) Decision Tree (DT)
3) Naïve Bayes (NB)
4) Linear Support Vector Machine (LSVM)
5) Polynomial Support Vector Machine (PSVM)

6) Radial Basis Function Support Vector Machine (RBF SVM)
7) K-Nearest Neighbors (KNN)
8) AdaBoost (AD)
9) Linear Discriminant Classifier (LD)
10) Quadratic Discriminant Classifier (QD)
11) Random Forest (RF)

In order to decide what evaluation metric to choose for the wrapper, we contacted an OB-GYN to determine what clinicians usually look for in medical applications. After a thorough discussion, it was clear that it is important to reduce the number of false positives (FP) so that patients are not put on a course of treatment that would do more harm than good. Thus, precision was chosen as the studied score since high precision, as shown in the Equation 2, means that a low number of FP was achieved.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Where TP and FP and FN are the number of true positives and false negatives correspondingly [17]. An example of the results obtained after implementing the SFFS algorithm can be seen in Figure 3.

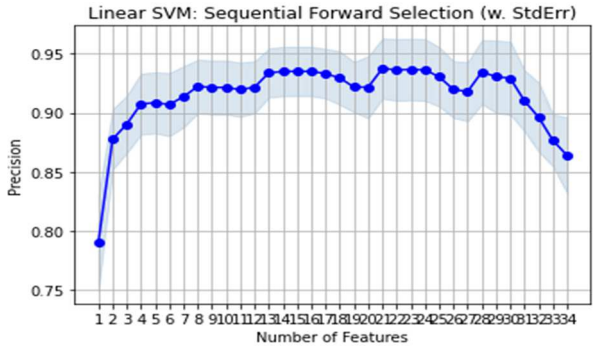

Figure 3. SFFS Wrapper Applied on Linear SVM to Evaluate its Precision

In order to determine which classifier performed best when it comes to precision, we compared each two classifiers using the paired T-test on the obtained results. Table III shows some of the p-values that were obtained after applying the paired T-test.

TABLE III. P-VALUES FOR EACH SET OF CLASSIFIERS BASED ON THE PRECISION VALUES

Classifier	LSVM	PSVM	RBF SVM
LSVM		**0.0482 QD better than LSVM**	0.0502
PSVM	**0.0378 PSVM better than LSVM**		**0.0033 RBF SVM better than PSVM**
RBF SVM	0.0502	**0.0033 RBF SVM better than PSVM**	
KNN	0.0646	0.3434	**0.0028 KNN better than RBF SVM**

After analyzing the results, it was concluded that the best performing classifiers were the KNN, Linear SVM, and the Polynomial SVM. However, the results of the t-test for these three classifiers were inconclusive and the null hypothesis could not be accepted or rejected.

Therefore, in order to narrow down which classifier performed best, the following steps were taken.

- For each model, the subset of features that yielded the highest precision was taken
- For each classifier, its corresponding precision-based selected subset was reintroduced into it
- The model was evaluated in terms of accuracy and recall as shown in Table IV
- The t-test was applied to the accuracy and the recall; the p-value was computed for each pair of the classifiers (Table V)

TABLE IV. DIFFERENT EVALUATION METRICS VALUES AFTER INTRODUCING THE BEST PRECISION FEATURE VECTORS INTO THE SELECTED CLASSIFIERS

Metric \ Classifier	Linear SVM	Polynomial SVM	KNN
Precision	0.936645	1	0.990909
Recall	0.8066	0.295221	0.573897
Accuracy	0.916	0.7697	0.8067

TABLE V. P-VALUES FOR EACH SET OF CLASSIFIERS BASED ON THE ACCURACY AND RECALL METRICS

Metric \ Combination	LSVM - PSVM	PSVM – KNN	KNN – LSVM
Recall	0.000001	2.71E-05	0.000204
Accuracy	0.014389	0.06221	0.096515

Based on the obtained results the Linear SVM was chosen as the optimal model since it had the best performance when it came to all three metrics. As for the final feature vector, the following 24 attributes were considered: Age, Weight, Height, BMI, Pulse Rate, Hb, Cycle Length, Marriage Status, Hip. Waist, TSH, PRL, PRG, RBS, Follicle No. (L), Follicle No. (R), Endometrium, Weight gain, Hair growth, Hair loss, Pimples, Fast food, and Regular Exercise.

D. Training and Evaluating Model

To avoid biased data splitting, K-Fold Cross Validation (KCV) technique, which is one of the most effective and reliable resampling techniques [18], was implemented.

Our model was evaluated, after feature reduction, by measuring the precision, accuracy, and recall metrics. After introducing the best feature vector to the LSVM and applying 10-fold cross-validation to train and test the classifier, the model achieved an accuracy of 91.6%, a precision of 93.66%, and a recall of 80.6%.

IV. RESULTS AND DISCUSSION

After comparing our results to the research found in the literature, it was concluded that our proposed model utilized medically reliant features and had high precision performance when testing on a high number of subjects. To illustrate, all previous models either did not use all possible

medical features or tested their results on a low number of subjects using hold-out method (Table VI). In our case, on the other hand, we were able to achieve high results on a high number of patients while taking into consideration reliable features.

TABLE VI. COMPARISON BETWEEN OUR PROPOSED MODEL AND OTHER RESEARCH WORK

Work	Features	No of Subjects	Evaluation Method	Accuracy %	Precision %
Mehrotra et al	Clinical & metabolic	200	Hold out	91.04	82.5
Vikas et al	Lifestyle & food intake	119	Hold out	97.65	95
Denny et al.	Clinical metabolic imaging hormonal & biochemical	541	Hold out	89	95.83
Proposed method	Clinical, metabolic imaging, hormonal & biochemical	541	10-Fold CV	91.6	93.66

V. CONCLUSIONS & FUTURE PERSPECTIVES

In this paper, we have investigated the use of classification algorithms to identify patients with Polycystic Ovary Syndrome (PCOS). We studied several machine learning algorithms on a large dataset using cross-validation technique, statistical tests, and physician opinion. Our objective was to maximize the precision of the model based on our discussion with physicians. In order to enhance the performance of our model, we used Sequential Forward Floating Selection (SFFS) to select the best features. The best algorithm was found to be Linear Support Vector Machine with 24 features. Both precision and accuracy of this algorithm were above 90%, whereas its recall was around 80%.

After realizing the importance of automating the process of PCOS diagnosis, and after reaching promising results in terms of precision, it is clearly seen that it is crucial to move forward with this research, and find the right approach to enhance the model's performance. Our analysis of the results showed that the classifiers had relatively high precision but did not perform as well when it came to recall. Therefore, there was a need to find a tradeoff between precision, accuracy, and recall. For that, we propose proceeding with the research as below:

- Relying on accuracy as the evaluation metric of the proposed model
- Applying the SFFS wrapper to the above classifiers to improve the accuracy metric
- Computing the precision and recall for each classifier when given their own optimal set of features (according to accuracy)
- Choosing the classifier with the optimal performance with respect to the three metrics through method of elimination
- Building our own database with a larger number of subjects

REFERNCES

[1] S. Palomba, Infertility in Women with Polycystic Ovary Syndrome, Springer, 2018.

[2] A. Denny, A. Raj, A. Ashok, C. Ram and R. George, "i-HOPE: Detection And Prediction System For Polycystic Ovary Syndrome (PCOS) Using Machine Learning Techniques," in *IEEE Region 10 International Conference TENCON*, 2019.

[3] H. Vassalou, M. Sotiraki and L. Michala, "PCOS diagnosis in adolescents: the timeline of a controversy in a systematic review," *J Pediatr Endocrinol Metab,* vol. 32, no. 6, pp. 549-559, 2019.

[4] C. Dennett and J. Simon, "The Role of Polycystic Ovary Syndrome in Reproductive and Metabolic Health: Overview and Approaches for Treatment," *Diabetes Spectrum,* vol. 28, no. 2, pp. 116-120, 2015.

[5] T. Williams, R. Mortada and S. Porter, "Diagnosis and Treatment of Polycystic Ovary Syndrome," *Am Fam Physician,* vol. 94, no. 2, pp. 106-113, 2016.

[6] R. Madhavan, "Prevalence of PCOS diagnoses among women with menstrual irregularity in a diverse, multiethnic cohort," 2015.

[7] T. Copp, E. Cvejic, K. McCaffery, J. Hersch, J. Doust, B. Mol, A. Dokras, G. Mishra and J. Jansen, "Impact of a diagnosis of polycystic ovary syndrome on diet, physical activity and contraceptive use in young women: findings from the Australian Longitudinal Study of Women's Health," *Human Reproduction,* vol. 35, no. 2, p. 394–403, 2020.

[8] P. Mehrotra, J. Chatterjee, C. Chakraborty, G. Biswanath and S. Ghoshdastidar, "Automated screening of Polycystic Ovary Syndrome using machine learning techniques," in *2011 Annual IEEE India Conference*, 2011.

[9] B. Vikas, B. Anuhya, M. Chilla and S. Sarangi, "A Critical Study of Polycystic Ovarian Syndrome (PCOS) Classification Techniques," *International Journal of Clinical and Experimental Medicine,* vol. 21, no. 4, pp. 1-7, 2018.

[10] P. Kottarathil, *Polycystic ovary syndrome (PCOS) Dataset,* Kaggle, 2020.

[11] A. O'Malley, K. Draper, R. Gourevitch, D. Cross and S. Scholle, "Electronic health records and support for primary care teamwork," *J Am Med Inform Assoc,* vol. 22, no. 2, p. 426–434, 2015.

[12] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing,* vol. 97, 2019.

[13] S. Solorio-Fernández, J. Martínez-Trinidad, J. Carrasco-Ochoa and Y.-Q. Zhang, "Hybrid feature selection method for biomedical datasets," in *IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology*, 2012.

[14] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Comput Biol Med,* vol. 112, pp. 1-9, 2019.

[15] E. Ostertagová and O. Ostertag, "Methodology and Application of Oneway ANOVA," *American Journal of Mechanical Engineering,* vol. 1, no. 7, pp. 256-261, 2013.

[16] Y. Wah, N. Ibrahim, H. Abdul Hamid, S. Abdul-Rahman and S. Fong, "Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy," *Pertanika Journal of Science and Technology,* vol. 26, no. 1, pp. 329-340, 2018.

[17] M. Marino, Y. Li, M. Rueschman, J. Winkelman, J. Ellenbogen, J. Solet, H. Dulin, L. Berkman and O. Buxton, "Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography," *Sleep,* vol. 36, no. 11, pp. 1747-1755, 2013.

[18] D. Anguita, A. Ghio, S. Ridella and D. Sterpi, "K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines," in *The 2009 International Conference on Data Mining*, Las Vegas, USA, 2009.