Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2022) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022)

Evaluation of Machine Learning Techniques for Improving Interpretation and Diagnostic Accuracy of Cardiopulmonary Exercise Training Data

Adrian Santos[a], Joackin Santos[a], Kyle Aquino[a], Vincent Magboo[b], Sheila Magboo[c]

[a]*Student, Zone 72, 670 Padre Faura St, Ermita, Manila, 1000 Metro Manila, Philippines*
[b]*Professor, Zone 72, 670 Padre Faura St, Ermita, Manila, 1000 Metro Manila, Philippines*
[c]*Professor, Zone 72, 670 Padre Faura St, Ermita, Manila, 1000 Metro Manila, Philippines*

Abstract

Cardiopulmonary Exercise Testing (CPET) is a diagnostic tool used in identifying limitations in the heart, lungs, and other such bodily parts or systems that are crucial to sustaining exercise. Evaluation of CPET data is useful as an alternative to more expensive and invasive diagnostic procedures, but for decades has remained underutilized for the complexity of its data leading to contradicting diagnosis by even trained clinicians. For this reason, this paper explored the viability of machine learning to act as alternative to traditional methods of interpreting CPET data such as flowcharts, and thus measured the effectiveness of six machine learning algorithms, namely: Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, and Autoencoder with Logistic Regression, in identifying the limitation afflicting each case in a CPET dataset, by way of their accuracy, precision, recall, and f1-score. F1-score was chosen as the sole metric to base the effectiveness or predictive power of a model because of the imbalanced nature of the dataset, and the metrics of which were further compared to a baseline model Logistic Regression, which was chosen for its simplicity relative to the aforementioned models.

Each of the seven models, including the baseline model, were tested in six cases involving the dataset differing based on feature selection and resampling, and were reran 1000x to get their average metric scores. The results indicate that Support Vector Machine performed best overall with an f1-score of 0.69 when there's resampling but there's no feature selection, it has also outperformed the baseline model Logistic Regression with its f1-score of 0.63 under the same conditions regarding the dataset. The models' performance under the six different dataset cases also indicate that models perform best with resampling but with no feature selection, indicating that all 100+ features of the dataset are important and contribute to the data's classification.

Keywords: CPET ; Machine Learning ; Neural Networks

1. Introduction

Cardiopulmonary Exercise Testing (CPET) is a diagnostic tool used to assess limitations during exercise through breath-by-breath gas exchange analysis. In particular, it is useful in identifying limitations in bodily parts that are crucial to sustaining exercise, which are the lungs, heart, circulatory system, and the aspect of the skeletal muscle for O2-CO2 exchange for aerobic mitochondrial metabolism. CPET data however consists of a multitude of time series data that prove difficult to interpret, even for trained and experienced clinicians, resulting in varying interpretations of the same results and CPET's subsequent underutilization [1]. For more than 20 years, various methods have been proposed to tackle its interpretability, including a flowchart consisting of a branching series of questions for easier diagnosis, but even that led to contradicting diagnosis by different clinicians and CPET remained underutilized [2]. For this, various papers have come to explore the possible use of machine learning models to both enhance the diagnostic power and interpretability of CPET data, many of whom have also been encumbered by the lack of available CPET data. This paper is one of them, and in this study, we will apply various machine learning algorithms in a bid to explore effective models in both interpreting and enhancing the accuracy of diagnosis using CPET data. In particular, to identify and differentiate between patients experiencing pulmonary, cardiac, muscle-skeletal limitations, or are otherwise healthy, while aiming for an f1-score and other such metrics that prove the model to be above what baseline models could do.

Nomenclature

VO2 (Oxygen consumption)	Rate of oxygen consumption expressed in absolute terms (mL/min), (L/min) or relative (mL/kg/min).
VO2 peak (Peak oxygen consumption)	Greatest rate of oxygen consumption during maximal progressive exercise; a measure of cardiorespiratory fitness.
VCO2 (Carbon dioxide production)	Rate of carbon dioxide exhaled (mL/min).
HR (Heart Rate)	Number of beats per minute (bpm); a variable progressing from less than 100 bpm while resting and increasing during progressive exercise to a peak.
VE (Minute ventilation)	Volume of air exhaled per minute (L/min).
RER (Respiratory exchange ratio)	Molar ratio of CO2 produced per O2 consumed; a variable progressing from less than 0.80 to greater than 1.10 during progressive exercise.
RR (Respiratory Rate)	Number of breaths per minute.
O2 pulse (Oxygen pulse)	Volume of oxygen uptake per heartbeat (VO2/HR; mL/beat); an alternative measure for stroke volume.
VE/VCO2 (Slope of minute ventilation versus carbon dioxide product)	Measurement of ventilatory efficiency or dead space ventilation.
VT (Ventilatory threshold)	Point in time when ventilation disproportionally increases compared to oxygen consumption (VO2 @ VT (mL/min) or HR at VT (beats/min)); reflects increased energy demands from anaerobic metabolism.

2. Literature Review

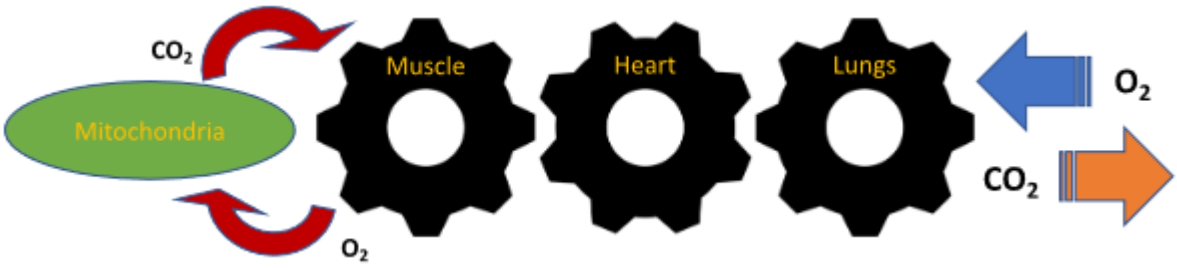


Fig. 1. Coupling of external to cellular respiration. Taken from Brown et al. (2022)'s adaptation from Wasserman K. Am J Physiol.1994 Apr;266(4 Pt 1):ES19-39.

Portella et al. (2022) characterized CPET data as unique in evaluating human response to maximal exercise stress, and while remarking it as useful in identifying specific limitations to exercise and as an alternative to more expensive and invasive diagnostic tests, noted that CPET data still remains underutilized for the complexity of its data making it hard to interpret. Their study sought to explore the use of machine learning models to aid in both interpretation and visualization of test results, for which they used Automated Machine Learning (AutoML) to automate the process of finding the optimum machine learning model with which to classify CPET data. Of the more than 1000 models their study has explored through AutoML, random forest was found to be the best machine learning model for classifying their CPET data with a mean accuracy of 0.866 and an Area under the ROC Curve (AUC) mean score of 0.898. Moreover, in the process of data gathering, Portella's team gathered data from 225 CPET cases, engineered key features, and thus created the Processed CPET dataset that is used in this study. [1]

In a separate study, Brown et al. (2022) characterized CPET data on a similar note and added that there have been attempts to increase the interpretability of CPET data both with and without machine learning. Flowcharts are one such endeavor, and have already been around for over 20 years, but had little effect in both interpretability and diagnostic capability as even trained clinicians struggled with contradicting results over the same data. Their study explored the feasibility of using deep learning neural networks to classify CPET data, where CPET data is generally scarce due to its underutilization and where such neural networks typically require huge amounts of data. The study has found the autoencoder with logistic regression model to be most effective with an average accuracy of 94% as compared to baseline models, which are the flowchart and principal component analysis (PCA) with logistic regression methods, with average accuracy scores at 77% and 90% respectively.

Andonian et al. (2021) presented a standardized methodology for interpreting cardiopulmonary exercise test (CPET) data, which is commonly utilized to evaluate exercise limitations in various physiological systems, including the cardiovascular, pulmonary, and skeletal muscle systems. Specifically, cardiac limitations can be inferred through an analysis of the shape of the O2-pulse curve, as well as the slope of the VE/VCO2 ratio. Pulmonary limitations can be identified by evaluating indices such as the VE/MVV ratio and the VE/VCO2 slope. Limitations in skeletal muscle function can be detected through measurements of VT, VE/VO2, and the rate of change in cardiac output (d(CO)/dt) during exercise [5]. Additionally, James et al. (2020) identified predictors of Chronic Obstructive Pulmonary Disease (COPD) using cardiopulmonary exercise test (CPET) data, which is a chronic lung condition known to lead to multiple exercise limitations, including pulmonary, skeletal muscle, and cardiac. They found that the dominant abnormalities in patients with COPD can be detected through the analysis of VE/VCO2 measurements [6].

The potential of using computer-aided algorithms, specifically machine learning techniques, used in the field of respiratory diseases is of interest. These algorithms were tested to evaluate CPET data in this study to identify chronic heart failure (CHF), and chronic obstructive pulmonary disease (COPD). The data set was collected from

234 CPET files from two medical centers based in Israel, classified as having confirmed CHF, COPD, and healthy individuals. This data set was used to train and validate a support vector machine (SVM) model. The overall predictive power of the proposed model demonstrates to be from 96% to 100% with significantly high scores of sensitivity, specificity, and precision. Therefore, the study concluded that its proposed model is capable of classifying patients to CHF, COPD, or healthy classes and capable for clinical applications [4].

The ERS task force, which comprises respiratory professionals, conducted a thorough review of existing literature and guidelines for standardizing CPET in patients with chronic lung diseases. Upon reviewing, they were able to determine current guidelines and recommendations for this objective. One is that proper patient preparation, through the use of bronchodilators and corticosteroids, is significant in having accurate test results. Moreover, it was also emphasized that careful interpretation of test results is necessary relative to the diseases subject to evaluation. The task force emphasized on using consistent and standardized protocols when conducting CPET to ensure that results are common and comprehensive for others to use, as well as ensure accurate diagnosis and effective treatment plans for the patients [3].

3. Methodology

The machine learning methodology employed in this study consisted of several steps. First, we collected and preprocessed the data, which included data cleaning, dataset normalization, selecting relevant features, and addressing any imbalances in the dataset. Next, we applied a variety of machine learning models to the data and evaluated their performance using a hold-out test set and a variety of evaluation metrics. We selected the model with the best performance for further analysis. The complete machine learning pipeline is depicted in Figure 2.

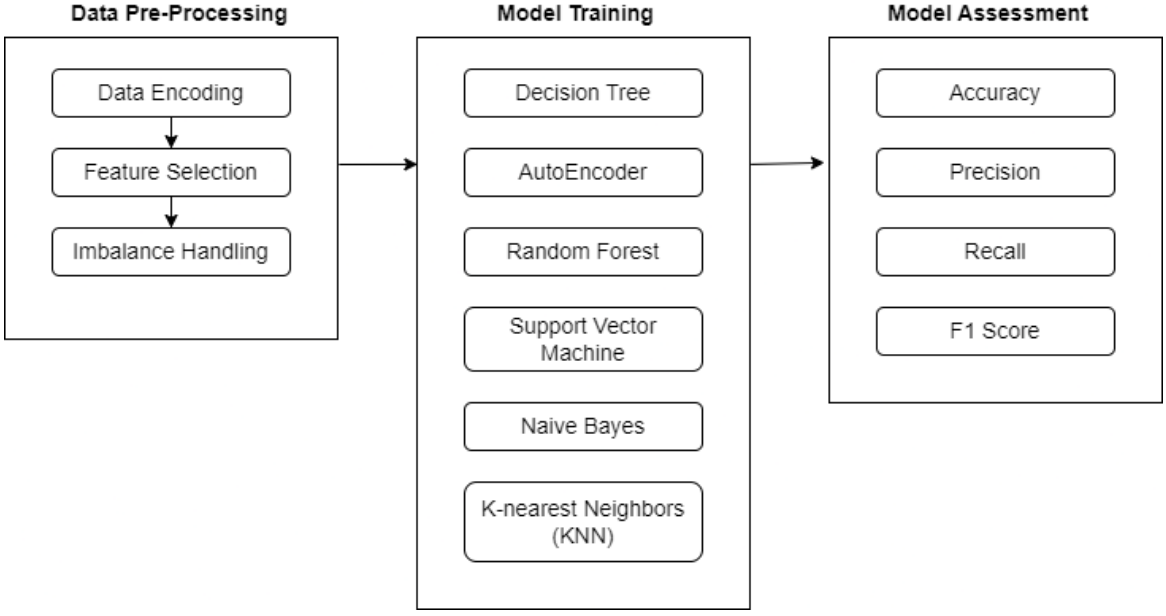


Fig. 2. Machine Learning Pipeline

3.1. Dataset Description

For this study, we used the Processed Cardiopulmonary Exercise Test (CPET) dataset taken from the works of Portella et al. (2022), a dataset publicly available through IEEE which has already been preprocessed and contains over 100 features and 4 labels. These features are time series data taken from different measurements of the

following, as per the definition in the aforementioned study of Portella et al.'s and are defined in this study's nomenclature.

The labels, which are binary, indicate whether the patient is healthy or has primary cardiac, pulmonary, or muscle skeletal limitations. The dataset contains 217 instances which is composed of 95 healthy, 51 cardiac limitation, 45 muscle skeletal limitation, 26 pulmonary limitation. This distribution is depicted in Figure 3.

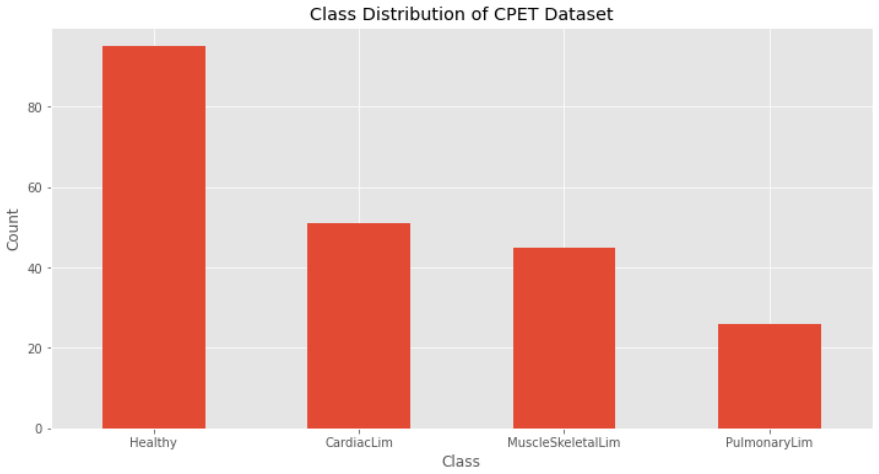


Fig. 3. Class Distribution of CPET dataset

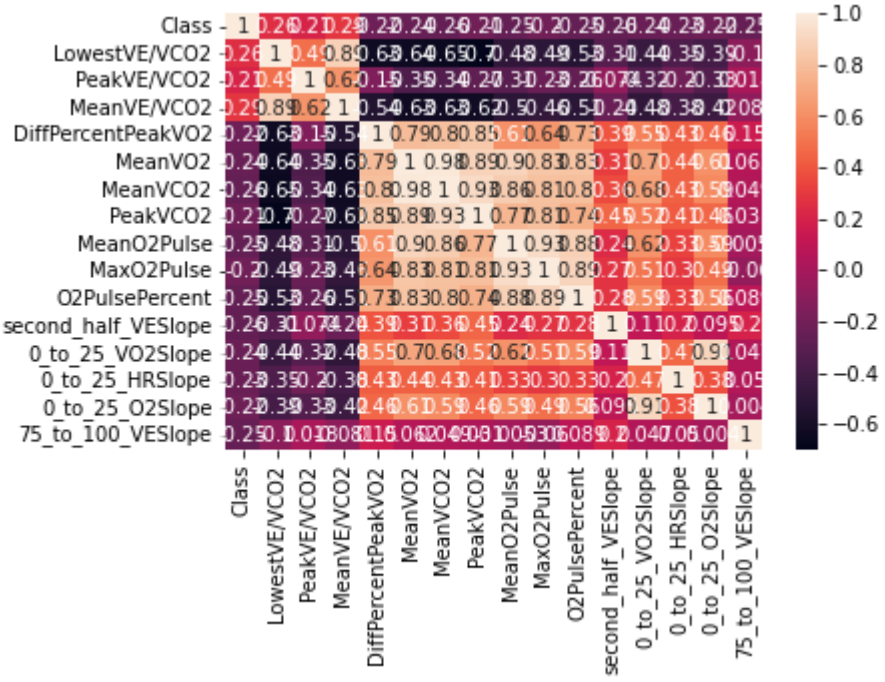
3.2. Pre-processing Steps

In this study, we applied a variety of preprocessing methods to the dataset in preparation for machine learning (ML) training. First, we checked for any missing or invalid values and duplicate records and found that there were none. Next, we assessed the balance of the dataset and found that there was a mild imbalance with 95 (43.78%) of samples belonging to healthy, 51 (23.50%) belonging to cardiac limitation, 45 (20.73%) belonging to muscle skeletal limitation, and 26 (11.90%) belonging to pulmonary limitation.

To prepare the data for ML algorithms, we performed data encoding for categorical attributes and feature scaling using normalization. Specifically, we applied the StandardScaler function using the scikit-learn library. Additionally, we dummified all categorical predictors, resulting in an increase in the number of columns.

To select the most relevant features for our model, we used two filter methods: Pearson Correlation and Mutual Information. We set a threshold correlation with the target variable of > 0.2 and acquired the top 10 information gain values of features with respect to the target variable.

Because of the large gap between the "Healthy" and "PulmonaryLim" classes standing at 67:21 ratio, Synthetic Minority Over-samping Technique (SMOTE) and Random Under Sampler were used to produce synthetic records for the minority classes and reduce the size of the majority classes respectively, making the 4 classes match with 34 records each. The results of the feature selection using Pearson correlation, information gain, and their correlation heatmaps can be seen in Fig. 4 and Fig. 5 respectively.



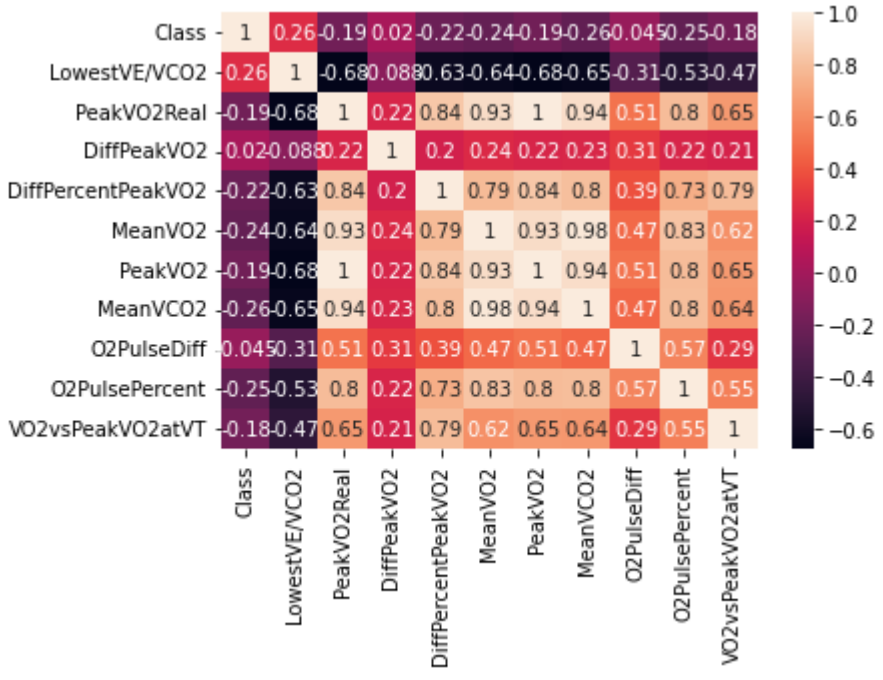


Fig. 4. Correlation Heatmap of Predictor Variables and Target Variable (Pearson Correlation)

Fig. 5. Correlation Heatmap of Predictor Variables and Target Variable (Information Gain)

3.3. Machine Learning Models

The preprocessed dataset was partitioned into a test set (30% of the data) and a training set (70% of the data). The training set contained 268 observations while the test set contained 66 observations. We utilized Python 3.8 and several machine learning libraries, including pandas, numpy, matplotlib, imblearn, and scikit-learn (sklearn), to implement and evaluate a range of ML models, including: Naive Bayes, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, and Autoencoder with Logistic Regression. Hyperparameter tuning through GridSearchCV was conducted for each model to optimize their performance. The F1 score was used to determine the best performing model, while each of the aforementioned models were also separately compared against a baseline statistical model Logistic Regression.

The models investigated are described as follows:

- **Naive Bayes.** A probabilistic machine learning algorithm that uses Bayes' theorem to make predictions. It makes the assumption that all features are independent given the class label, allowing for simple and efficient computation of the class label given the feature values. It is commonly used for classification tasks [10].
- **Decision Tree.** An easily visualizable machine supervised machine learning algorithm that builds a so-called decision tree in a top-down manner, each time splitting or partitioning the data based on the input features in such a way that the resulting leaves contain data points that are similar to each other with respect to the class. [8]
- **Random Forest.** This is an ensemble machine learning algorithm that creates a set of decision trees (called a forest) from a randomly selected subset of training data. Afterwhich, the output of the set of decision trees is averaged or selected by majority vote to obtain a final output, otherwise known as bagging. This type of approach reduces overfitting in the model. This algorithm is useful for both classification and regression tasks, and has the ability to handle high dimensional data with numerous features. Among the ensemble techniques, it is considered one of the most accurate and robust algorithms [13].
- **K-Nearest Neighbors.** A supervised machine learning algorithm that can be used for both classification and regression tasks. It makes predictions based on the majority class or average value of the K nearest data points to a given test point in the feature space. The value of K is a user-specified parameter, with a smaller K value leading to a more complex and potentially overfitting model, and a larger K value leading to a smoother and potentially underfitting model. It is simple to implement and does not require any assumptions about the underlying data distribution [11].
- **Support Vector Machine.** A type of supervised machine learning algorithm used for classification and regression tasks. It works through finding the hyperplane in a high-dimensional space that maximizes the separation of different classes. Moreover, it introduces support vectors which are points closest to the hyperplane that determines the position of the hyperplane. This algorithm is useful for cases when the number of features are greater than the number of samples, and when data is non-linearly separable using kernels [12].
- **Logistic Regression.** A supervised machine learning algorithm that is typically used to fit a regression model for predicting the response variable, the class, is binary (categorical). The model can be extended to multi-class classification, and is here used as the baseline model for being relatively simpler compared to the other models mentioned above. [7]
- **Autoencoder with Logistic Regression.** A type of unsupervised neural network consisting of encoder and decoder networks, that can be configured to be deeply nested, that maps data to their lower-dimension representations and back respectively. The idea is to train the autoencoder model to effectively reconstruct the input data while minimizing the difference between the input data and the output of the decoder network. Autoencoders are thus often used for dimensionality reduction, denoising, and the output of which here is fed to a logistic regression model for classification of the data. [9]

4. Results

Each of the aforementioned models, including the baseline Logistic Regression model, were measured through the following metrics: accuracy, precision, recall, and f1-score. Accuracy is the ratio of correct predictions to the total number of predictions made, precision is the ratio of the number of patients whose limitation was correctly classified to the total number of such predictions, recall, or sensitivity, is the ratio of the same correctly classified patients to the number of patients that actually have the limitation, while f1-score is the harmonic mean or trade-off between precision and recall. F1-score is used as the sole deciding factor for the performance of the model as it performs well even in imbalanced datasets where accuracy may have otherwise painted misleading results.

Hyperparameter tuning was done using GridSearchCV, where the resulting model configuration is then run 1000x. Each time, the accuracy, precision, recall, and f1-score for the iteration was taken, where the averages of which were taken after the 1000x iterations have elapsed. The average scores for the four metrics are illustrated in the tables below.

Table 1. Performance Metrics for Dataset without resampling.

	ML Model	Accuracy	Precision	Recall	F1-Score
No Feature Selection	LR	0.67	0.73	0.67	0.64
	NB	0.61	0.62	0.61	0.59
	DT	0.53	0.56	0.53	0.54
	RF	0.62	0.65	0.62	0.59
	KNN	0.68	0.69	0.68	0.66
	SVM	0.65	0.66	0.65	0.62
	AutoE + LR	0.67	0.73	0.67	0.64
Pearson's Correlation	LR	0.58	0.53	0.58	0.50
	NB	0.58	0.62	0.58	0.57
	DT	0.62	0.63	0.62	0.61
	RF	0.61	0.64	0.61	0.60
	KNN	0.58	0.54	0.58	0.55
	SVM	0.61	0.61	0.61	0.60
	AutoE + LR	0.58	0.53	0.58	0.50
Mutual IG (Information Gain)	LR	0.56	0.73	0.56	0.50
	NB	0.58	0.58	0.58	0.57
	DT	0.61	0.68	0.61	0.62

	RF	0.58	0.56	0.58	0.56
	KNN	0.61	0.47	0.48	0.47
	SVM	0.62	0.68	0.62	0.61
	AutoE + LR	0.56	0.73	0.56	0.50

Table 1 displays the performance assessment of seven machine learning models on various feature sets, evaluated on the dataset without resampling. The evaluation was carried out using several standard metrics such as accuracy, precision, recall, and F1-score. The results indicate that among the models, the K-Nearest Neighbors (KNN) algorithm achieved the highest F1-score of 0.66 when no feature selection was applied. However, when feature selection was performed using Pearson's Correlation and Mutual Information Gain, the Decision Tree algorithm achieved the highest F1-score of 0.61 and 0.62, respectively. In conclusion, the results suggest that the KNN algorithm demonstrated the highest overall performance among the evaluated models.

Table 2. Performance Metrics for Dataset with resampling.

	ML Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
	LR	0.62	0.67	0.62	0.63
No Feature	NB	0.58	0.65	0.58	0.57
Selection	DT	0.58	0.51	0.58	0.52
[Oversampling	RF	0.61	0.65	0.61	0.61
+ Undersampling]	KNN	0.59	0.69	0.59	0.60
	SVM	0.68	0.70	0.68	0.69
	AutoE + LR	0.62	0.70	0.62	0.63
	LR	0.64	0.70	0.64	0.64
Pearson’s	NB	0.53	0.57	0.53	0.53
Correlation	DT	0.55	0.66	0.55	0.58
[Oversampling	RF	0.61	0.64	0.61	0.61
+ Undersampling]	KNN	0.58	0.59	0.58	0.57
	SVM	0.65	0.66	0.65	0.65
	AutoE + LR	0.62	0.67	0.62	0.62
	LR	0.62	0.67	0.62	0.62
Mutual IG	NB	0.55	0.56	0.55	0.54

(Information Gain)	DT	0.47	0.53	0.47	0.49
[Oversampling	RF	0.65	0.67	0.65	0.65
+ Undersampling]	KNN	0.64	0.66	0.64	0.64
	SVM	0.65	0.67	0.65	0.66
	AutoE + LR	0.62	0.73	0.62	0.62

Table 2 presents the evaluation results of seven machine learning models on various feature sets, evaluated using the dataset with resampling. The performance of the models was also quantitatively assessed using several standard metrics, including accuracy, precision, recall, and F1-score. The results indicate that the Support Vector Machine (SVM) model demonstrated superior performance across all feature sets, with the highest performance observed when no feature selection was applied, achieving an F1-score of 0.69. The feature set utilizing Mutual Information Gain (MIG) also achieved a high performance, with an F1-score of 0.66. Lastly, the feature set utilizing Pearson's Correlation (PC) achieved an F1-score of 0.65. Overall, these results suggest that the SVM model demonstrated the best performance among the evaluated models when the dataset was balanced through resampling.

Table 3. Effect of undersampling and oversampling on the performance of machine learning models.

	ML Model	**Without Undersampling and Oversampling**	**With Undersampling and Oversampling**	**Difference (%)**
No Feature Selection	LR	0.64	0.63	-1.59
	NB	0.59	0.57	-3.5
	DT	0.54	0.52	-3.84
	RF	0.59	0.61	3.39
	KNN	0.66	0.60	-10
	SVM	0.62	0.69	11.29
	AutoE + LR	0.64	0.63	-1.59
	Average			-0.83
Pearson's Correlation	LR	0.50	0.64	28
	NB	0.57	0.53	-7.55
	DT	0.61	0.58	-5.17
	RF	0.60	0.61	1.67
	KNN	0.55	0.57	3.64

	SVM	0.60	0.65	8.3
	AutoE + LR	0.50	0.62	24
	Average			7.56
Mutual IG (Information Gain)	LR	0.50	0.62	24
	NB	0.57	0.54	-5.56
	DT	0.62	0.49	-26.53
	RF	0.56	0.65	16.07
	KNN	0.47	0.64	36.17
	SVM	0.61	0.66	8.20
	AutoE + LR	0.50	0.62	24
	Average			10.91

Table 2 presents the impact of undersampling and oversampling on the performance of machine learning models, evaluated using several standard metrics such as accuracy, precision, recall, and F1-score. The results suggest that the effect of the resampling technique varies depending on the feature set used. When no feature selection was applied, the resampling technique improved the performance of 28.57% (2 out of 7) of the models, with an average effect of -0.83%. In contrast, when the feature set utilizing Pearson's Correlation was used, the resampling technique improved the performance of 71.43% (5 out of 7) of the models, with an average effect of 7.56%. Similarly, when the feature set utilizing Mutual Information Gain was used, the resampling technique improved the performance of 71.43% (5 out of 7) of the models, with an average effect of 10.91%.

The results of our experiments, presented in Table 1 and Table 2, indicate that different machine learning models perform differently when evaluated on various feature sets and with different preprocessing techniques. In Table 1, the K-Nearest Neighbors (KNN) algorithm achieved the highest F1-score of 0.66 when no feature selection was applied, evaluated on the dataset without resampling. In contrast, Table 2 shows that the SVM model demonstrated superior performance across all feature sets, with the highest performance observed when no feature selection was applied, achieving an F1-score of 0.69. The feature set utilizing Mutual Information Gain (MIG) also demonstrated high performance, with an F1-score of 0.66.

The results suggest that all attributes are important in predicting primary limitations of patients, as the performance of most models is better when using that feature set. Additionally, the use of undersampling and oversampling has mixed effects on the machine learning models.

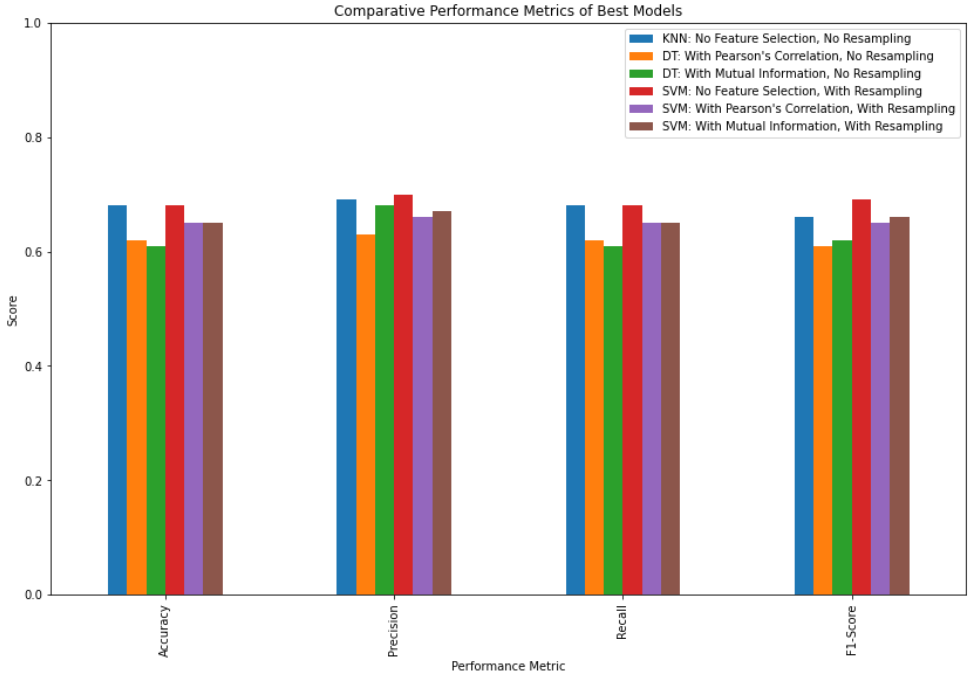


Fig. 6. Performance Metrics of Best Models

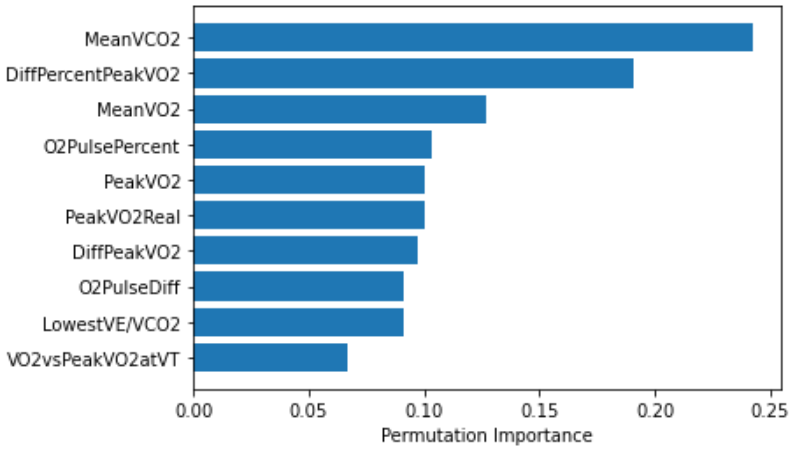


Fig. 7. Permutation Importance of SVM (Radial) model

Figure 6 represents the differences of each performance metric, namely: Accuracy, Precision, Recall and F1-score, for each of the top performing models based on whether or not resampling was employed on the training set, and if feature selection was performed. Overall, SVM without feature selection and with resampling has the highest scores. KNN without feature selection and no resampling ties with this model for accuracy and precision. However, it is important to note that these are not extreme cases and that the performance metrics are relatively close in score for the individual models. Therefore, it is crucial to consider the differences in model performance when feature selection is applied, to see which of the models are more desirable Upon comparing between the SVM models with resampling, and without feature selection and with feature selection, it can be observed that the difference in scores are miniscule, where using mutual information seems to be the most desirable.

Figure 7 meanwhile depicts the permutation importance of the features in terms of the SVM (Radial) model with resampling and with feature selection. Observed here, MeanVCO2 has the highest importance at approximately 0.24

and DiffPercentPeakVO2 as the second at approximately 0.18. These values show how much the prediction error increases when a feature is not available, therefore how important the feature is for the model in predicting.

Acknowledgements

The authors acknowledge the work of Portella et al. [1] for providing the data used in this study.

Appendix

The features features of the data used in this study were collected, engineered, and explained in this work of Portella et al. [1] https://doi.org/10.1109/JBHI.2022.3163402.

References

[1] J. Portella et al., “Using Machine Learning to Identify Organ System Specific Limitations to Exercise via Cardiopulmonary Exercise Testing,” IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 8. Institute of Electrical and Electronics Engineers (IEEE), pp. 4228–4237, Aug. 2022. doi: 10.1109/jbhi.2022.3163402.

[2] D. E. Brown, S. Sharma, J. A. Jablonski, and A. Weltman, “Neural network methods for diagnosing patient conditions from cardiopulmonary exercise testing data,” BioData Mining, vol. 15, no. 1. Springer Science and Business Media LLC, Aug. 13, 2022. doi: 10.1186/s13040-022-00299-6.

[3] T. Radtke, I. Vogiatzis, D. S. Urquhart, P. Laveneziana, R. Casaburi, and H. Hebestreit, “Standardisation of cardiopulmonary exercise testing in chronic lung diseases: summary of key findings from the ERS task force,” European Respiratory Journal, vol. 54, no. 6. European Respiratory Society (ERS), p. 1901441, Dec. 2019. doi: 10.1183/13993003.01441-2019.

[4] O. Inbar, O. Inbar, R. Reuveny, M. J. Segel, H. Greenspan, and M. Scheinowitz, “A Machine Learning Approach to the Interpretation of Cardiopulmonary Exercise Tests: Development and Validation,” Pulmonary Medicine, vol. 2021. Hindawi Limited, pp. 1–9, May 31, 2021. doi: 10.1155/2021/5516248.

[5] Andonian, B. J., Hardy, N., Bendelac, A., Polys, N., & Kraus, W. E. (2021). Making cardiopulmonary exercise testing interpretable for clinicians. Current Sports Medicine Reports, 20(10), 545–552. https://doi.org/10.1249/jsr.0000000000000895

[6] M. D. James, K. M. Milne, D. B. Phillips, J. A. Neder, and D. E. O’Donnell, “Dyspnea and Exercise Limitation in Mild COPD: The Value of CPET,” Frontiers in Medicine, vol. 7. Frontiers Media SA, Aug. 13, 2020. doi: 10.3389/fmed.2020.00442.

[7] S. Swaminathan, “Logistic regression - detailed overview,” Available at https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc (2018)

[8] W. Hillier, “What is a decision tree and how is it used?” Available at https://careerfoundry.com/en/blog/data-analytics/what-is-a-decision-tree/ (2021)

[9] J. Jordan, “Introduction to autoencoders,” Available at https://www.jeremyjordan.me/autoencoders/ (2018)

[10] K. El Hindi, H. AlSalman, S. Qasem, and S. Al Ahmadi, “Building an Ensemble of Fine-Tuned Naive Bayesian Classifiers for Text Classification,” Entropy, vol. 20, no. 11. MDPI AG, p. 857, Nov. 07, 2018. doi: 10.3390/e20110857.

[11] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao, and H. Yang, “A generalized mean distance-based k-nearest neighbor classifier,” Expert Systems with Applications, vol. 115. Elsevier BV, pp. 356–372, Jan. 2019. doi: 10.1016/j.eswa.2018.08.021.

[12] R. Gandhi, “Support Vector Machine - introduction to machine learning algorithms,” Medium, Jul. 05,2018. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[13] “What is Random Forest?,” IBM. https://www.ibm.com/topics/random-forest.