Leveraging Bio-Inspired Optimization Algorithms for Advanced Feature Selection in Chronic Disease Datasets

Abeer Dyoub[1,*,†], Ivan Letteri[2,†]

[1]Computer Science Department, University of Bari, Bari - Italy
[2]Department of Life, Health and Environmental Sciences,University of L'Aquila, L'Aquila - ITALY

Abstract

In this study, we investigated the application of bio-inspired optimization algorithms for feature selection in chronic disease prediction. The primary goal was to enhance models' predictive accuracy, streamline data dimensionality, and make predictions more interpretable and actionable. The research encompassed a comparative analysis of the three bio-inspired categories: evolutionary-based, swarm-intelligence, and ecology-based. For the feature selection method, we selected one algorithm for each category: Genetic Algorithms, Flower Pollination Optimization, and Particle Swarm Optimization, applying them across diverse chronic diseases including cancer, kidney, and cardiovascular diseases. The results demonstrate in some cases, that the bio-inspired optimization algorithms effectively reduce the number of features required for accurate classification and consequently the convergence time. The findings underscore this work's potential impact on early intervention, precision medicine, and improved patient outcomes, providing new avenues for delivering healthcare services tailored to individual needs.

Keywords

Chronic Diseases Prediction, Bio-Inspired Feature Selection, Genetic Algorithms, Flower Pollination Optimization, Particle Swarm Optimization

1. Introduction

Chronic diseases pose a significant global health challenge, impacting morbidity and mortality rates. Early detection is crucial for prevention and personalised healthcare. Advanced analytics and AI offer the potential for revolutionising prediction in many field like finance [1] [2], cybersecurity [3] and in particular disease.

Supervised learning in various fields relies heavily on feature selection (FS) to reduce input dimensionality. Maintaining target class integrity amidst irrelevant characteristics is essential for accurate classification in the medical domain.

Bio-inspired optimisation emulates behaviours found in various natural creatures such as fish, insects, bird swarms, terrestrial animals, reptiles, humans, and other phenomena. These methods have been used for supervised feature selection (see [4]). The same source categorises bio-inspired optimisation algorithms into three groups based on their source of inspiration: swarm intelligence algorithms, evolutionary-based algorithms, and ecology-based algorithms. For robustness and diversity, we selected Genetic Algorithms (GA), Particle Swarm Optimisation (PSO), and Flower Pollination Optimisation (FPO), one from each category.

We refine feature subsets from medical datasets encompassing cancer, kidney, and cardiovascular diseases to enhance model accuracy and simplify data dimensionality. The aim is to improve interpretability and practicality in chronic disease prediction.

Investigating chronic diseases presents significant challenges in the healthcare domain. This study aims to improve the predictive accuracy of chronic diseases by employing machine learning (ML) and feature selection (FS) techniques, which involve data collection, preprocessing, and performance assessment.

The paper proceeds with an outline of the methodology in Section 2. Section 3 presents experimental findings, followed by a discussion in Section 4. Finally, Section 5 summarises key findings, limitations, and future directions.

2. Methodology

Preprocessing techniques, including transformation, cleaning, imputation, balancing, and normalization, were applied to ensure data quality [5]. Subsequently, feature selection was performed by GA, PSO, and FPO algorithms. The selected features were then used for classification using Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), Support Vector Machines (SVM), and K-Nearest Neighbour (KNN). Finally, we evaluated the performance of these models using various metrics.

Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29-30, 2024, Naples, Italy
*Abeer Dyoub.
†These authors contributed equally.
abeer.dyoub@uniba.it (A. Dyoub); ivan.letteri@univaq.it (I. Letteri)
0000-0003-0329-2419 (A. Dyoub); 0000-0002-3843-386X (I. Letteri)

2.1. The Datasets

Breast Cancer dataset: From the University of Wisconsin, this dataset involves cytological examinations to distinguish between benign and malignant tumours. It contains 569 samples and 31 features.

Kidney Disease: Medical information on chronic kidney disease, collected over two months in India, is included in this dataset, available on Kaggle or UCI. It consists of 400 samples and 25 features.

Heart failure dataset: Comprising medical records of heart failure patients during follow-up, this dataset contains 299 samples and 13 features.

Each dataset has the "diagnosis" column with binary values used as targets for supervised learning of classifiers, where 0 denotes a negative and 1 indicates a positive outcome, respectively.

2.2. Datsets Pre-processing

Missing Values Imputation. Addressing missing data poses risks of performance degradation and biased results. We used the K-Nearest Neighbors (KNN) algorithm, known for its adaptability to diverse data types, to fill the lack in the datasets.

Data Balancing. To balance the datasets is a critical concern due to the struggle of the classifiers when faced with disparate class distributions, leading to biased models. To mitigate this issue, we used the Synthetic Minority Over-sampling Technique for Nominal and Continuous features (SMOTEEN) [6] which addresses imbalanced datasets by oversampling the minority class and cleaning the majority class by combining the SMOTE and Edited Nearest Neighbors (ENN) methods.

Min-max Normalization. We applied this scaling method to normalize the datasets to a predefined range, as follows: $X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$, where X_{norm} represents the normalized value of the feature, X is the original value of the feature. X_{min} and X_{max} denote the minimum and maximum values respectively.

2.3. Bio-inspired Feature Selection

Following the data preparation stage, we applied the three aforementioned bio-inspired feature selection algorithms to each of the three datasets (see section 2.1). All algorithms employ the same fitness function, with the α value set to 0.99 to prioritize classification accuracy.

For assessing fitness, we utilized the K-Nearest Neighbors (KNN) classifier, known for its efficiency and adopted by [7], as it does not necessitate a lengthy training phase. A neighbour count of $K = 10$ was used. The feature selection algorithms were configured with 20 agents (*individuals*) and 100 generations.

2.4. Performance Evaluation Method

For each dataset detailed in section 2.1, every machine learning model is trained using 70% of the data and tested using the remaining 30%, employing all features, and filtered features by PSO, FPO, and GA algorithms. This process is iterated 100 times with each iteration involving shuffling the dataset. Moreover, for each iteration, the dataset is split into training and testing sets to evaluate measures such as *Accuracy*, *Recall*, *Precision*, and *F1-score*.

3. Experiments and Results

Figures 1, 2, 3 show the fitness trends of the FS algorithms, and table 1 summarizes the performance of these FS algorithms in terms of feature reduction.

In table 2, we report the accuracies of the classifiers with the features. Whereas, in table 3, we report the percentage variations of the training time before and after the FS.

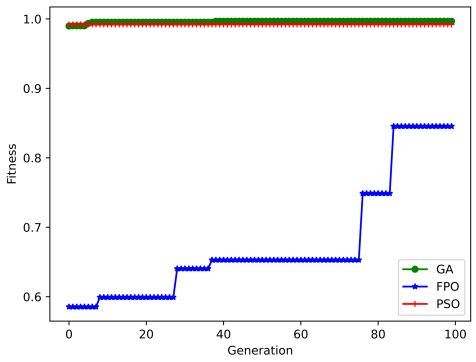

Figure 1: Fitness trends on breast cancer dataset

Table 1
Performance of Dimensional Reduction in the different Feature Selection Algorithms.

Dataset	Algorithm	Fitness	#Features	Reduction
Breast Cancer	GA	≈ 0.992	8	73.3%
	PSO	≈ 0.985	12	60%
	FPO	≈ 0.9092	8	73.3%
Heart Failure	GA	≈ 0.91	3	75%
	PSO	≈ 0.79	2	83.3%
	FPO	≈ 0.581	3	75%
Kidney Disease	GA	≈ 0.998	7	70%
	PSO	≈ 0.995	11	54%
	FPO	≈ 0.8454	5	80%

3.1. Breast Cancer Dataset

The final features selected by the various algorithms are:

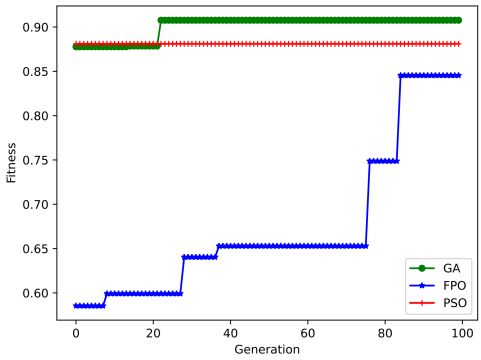

Figure 2: Fitness trends on heart failure dataset

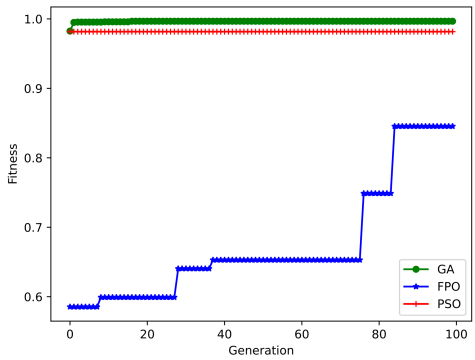

Figure 3: Fitness trends on kidney disease dataset

- *FPO*: ['sy11etry 1ean', 'fractal di1ension 1ean', 'radius se', 'area se', 'co1pactness se', 'sy11etry se', 'fractal di1ension se', 'concavity worst']
- *GA*: ['texture mean', 'concavity mean', 'area se', 'compactness se', 'concave points se', 'fractal dimension se', 'radius worst', 'compactness worst']
- *PSO*: ['radius mean', 'area mean', 'smoothness mean', 'compactness mean', 'fractal dimension mean', 'radius se', 'texture se', 'area se', 'smoothness se', 'compactness se', 'texture worst', 'concavity worst']

we note that GA has acheived a better fitness with respect to FPO, even both have achieved the same reduction percentage in dimentionality with breast cancer dataset. The two algorithms have selected different sets of features. Regarding the training time, we can observe that the training times have globally decreased up to a maximum of 54.5% for GA with LR on this dataset. In KNN, the time has increased in all the cases. further metrics can be seen in figures 4, 5, 6.

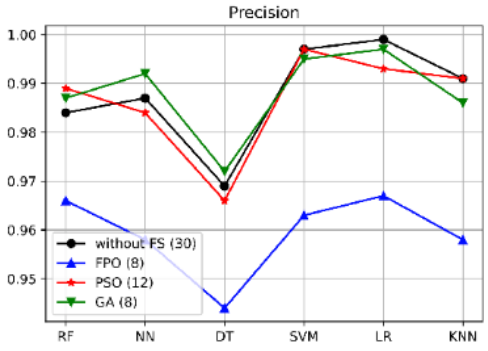

Figure 4: Precision on dataset Breast Cancer.

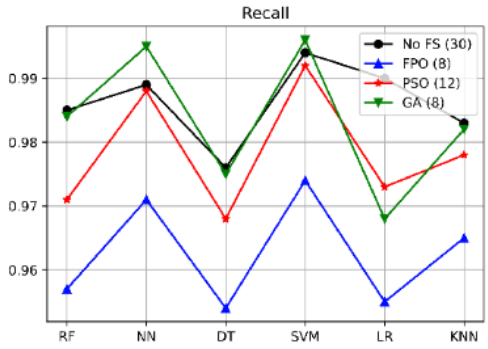

Figure 5: Recall on dataset Breast Cancer.

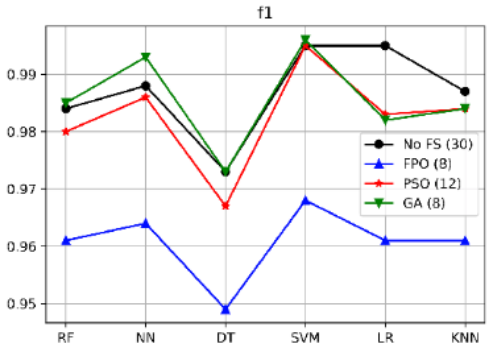

Figure 6: F1-score on dataset Breast Cancer.

3.2. Heart Failure Dataset

The final features selected by the three algorithms are:

- *FPO*: ['anaemia', 'diabetes', 'smoking']

- *GA*: ['platelets', 'serum sodium', 'time']
- *PSO*: ['platelets', 'serum creatinine']

we note that GA has acheived a better fitness with respect to FPO, even both have achieved the same reduction percentage in dimentionality with breast cancer dataset. The two algorithms have selected different sets of features. The genetic algorithm significantly higher fitness with respect to FPO and PSO even though the the dimentionality reduction is almost the same. RF, DT, SVM and KNN have achieved a better performance on this dataset when combined with GA algorithm. In general, however, the training times have all decreased, with the maximum decrease 57% by LR model with both GA and PSO. See further metrics in figures 7, 8 , 9.

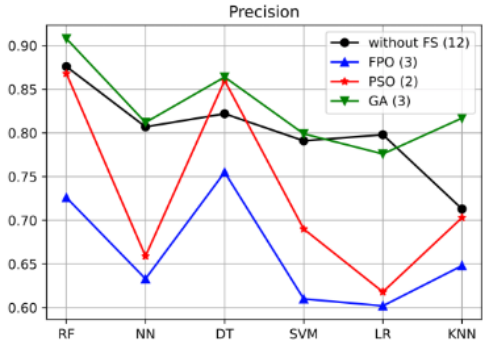

Figure 7: Precision on dataset Hearth Failure.

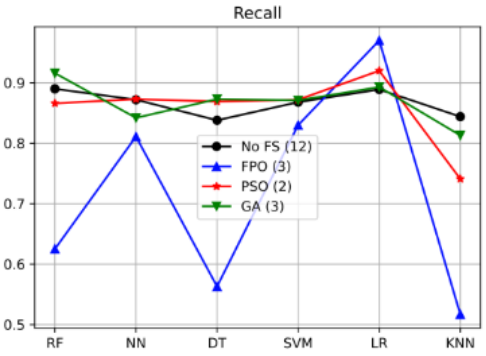

Figure 8: Recall on dataset Hearth Failure.

3.3. Kidney Disease Dataset

The final features selected by the various algorithms are:

- *FPO*: ['su', 'rbc', 'pcc', 'pe', 'ane']
- *GA*: ['rbc', 'bgr', 'sod', 'hemo', 'pcv', 'dm', 'cad']
- *PSO*: ['age', 'su', 'rbc', 'pc', 'bgr', 'sod', 'pot', 'hemo', 'pcv', 'rc', 'cad']

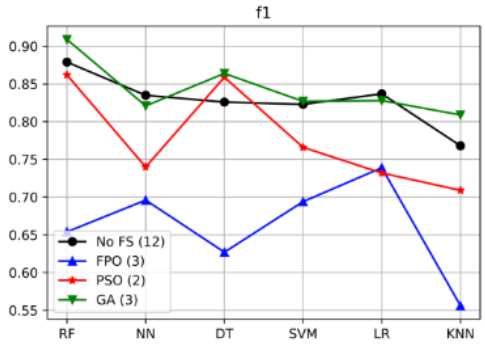

Figure 9: F1-score on dataset Hearth Failure.

In this dataset, high performance was achieved with most models combined with PSO and GA, while with FPO there was a significant decrease in performance. There has been a 55% decrease in processing time without loss in the performance with LR model combined with PSO and a reduction in processing time up to 57% with LR combined with GA with a very slight reduction in the performance. The highest fitness was achieved by GA with 7 features (70% reduction in dimentionality), while the lowest fitness was achieved by FPO with the highest features reduction. See further metrics in figures 10, 11, 12.

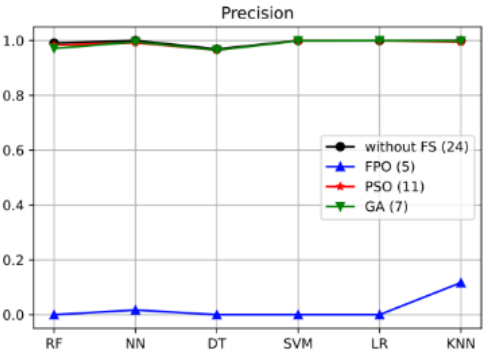

Figure 10: Precision on dataset Kidney Disease.

4. Discussion

From Table 2, it is evident that GA emerged as the most effective FS technique in terms of performance. It consistently improved accuracy across various ML models and datasets, or maintained accuracy levels compared to pre-FS values with other techniques. The accuracy enhancement with GA ranged from 0.1% to 7%. Following GA, PSO ranked second in terms of accuracy perfor-

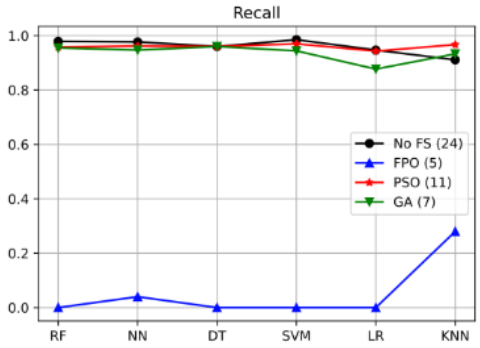

Figure 11: Recall on dataset Kidney Disease.

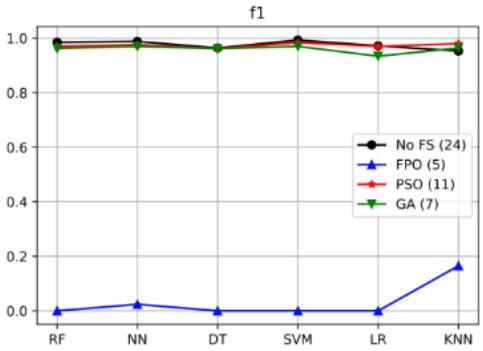

Figure 12: F1-score on dataset Kidney Disease.

mance among the three bio-inspired algorithms. While PSO did not notably enhance accuracy, it also did not lead to significant decreases. FPO exhibited diverse outcomes across different ML models and datasets. While accuracy decreases were marginal (less than 2.5%) for most ML models on the breast cancer dataset, there were more pronounced decreases on the heart failure and kidney disease datasets.

In terms of training times, the impact was particularly notable for DT and LR, as evidenced in Table 3. Generally, training times decreased across all models when employing feature selection (FS), except for K-Nearest Neighbours (KNN) with breast cancer and kidney disease datasets, where a significant increase of up to 21% was observed. Minor fluctuations within ±2% in training times were considered insignificant, likely due to variable hardware conditions and software factors. Overall, machine learning (ML) models exhibited reduced training times with FS, especially DT and LR models with GA and PSO. The most substantial reduction in training time, up to 67%, was achieved by the LR model with FPO on the Kidney disease dataset. Although FS did not significantly improve ML model performance in most cases, and even led to a decrease in performance in some instances, the noteworthy decrease in processing times without significant loss in accuracy represents a significant achievement.

The experimental findings indicate that the GA outperformed other FS algorithms in terms of precision, recall, and F1-measure. GA demonstrated superior performance when paired with nearly all ML models compared to FPO and PSO across all datasets. However, the PSO algorithm, when combined with the LR model, exhibited slightly higher recall and F1 scores for breast cancer and kidney disease datasets, as well as marginally improved recall for heart failure dataset. Conversely, FPO generally exhibited the poorest performance when paired with various ML models. Although FPO achieved the highest recall when combined with the LR model on the heart failure dataset, its overall performance was inferior. In terms of fitness trends, GA displayed the most favourable results, with PSO closely trailing behind, while FPO yielded significantly lower fitness levels compared to GA and PSO. Further experiments are planned to investigate the behaviour of these FS algorithms with varying parameters, datasets, and ML models.

Table 2
Models Accuracy.

		Breast Cancer	Heart Failure	Kidney Disease
RF	No FS	98.4%	85.5%	98.7%
	FPO	96.2%	61.2%	57%
	PSO	98%	83.6%	97.5%
	GA	**98.5**%	**89.2**%	96.8%
DT	No FS	97.3%	79.2%	96.9%
	FPO	94.9%	61.1%	57%
	PSO	96.7%	**83**%	96.8%
	GA	**97.3**%	**83.6**%	96.8%
SVM	No FS	99.5%	77.7%	99.4%
	FPO	96.9%	56.8%	57%
	PSO	**99.5**%	68.2%	98.7%
	GA	**99.6**%	**78.3**%	97.5%
LR	No FS	99.5%	79.4%	97.7%
	FPO	96.1%	59.1%	57%
	PSO	98.3%	59.9%	97.6%
	GA	98.3%	77.8%	94.7%
KNN	No FS	98.7%	69.6%	96.1%
	FPO	96.1%	52.9%	52.5%
	PSO	98.5%	64.6%	**98.4**%
	GA	98.4%	**77.3**%	**97.1**%

5. Conclusion

Our experiments have highlighted the importance of feature selection (FS) in improving the performance of machine learning (ML) models. The impact of FS varies depending on factors such as the chosen FS algorithm and dataset characteristics [8]. FS holds the potential to significantly enhance ML outcomes, especially for datasets with a large number of features. For example, in the breast cancer dataset, reducing features from 30 to 12 or 8 resulted in up to a 50% reduction in training time, while maintaining the same performance across various ML models. However, the effect of FS on training time

Table 3
Models Processing Time

		Breast Cancer	Heart Failure	Kidney Disease
RF	FPO	-7%	+2%	-2%
	PSO	-3.5%	+1%	-2%
	GA	-5%	+1%	-4%
DT	FPO	-40%	-8%	-25%
	PSO	-40%	-6%	-11%
	GA	-50%	-8%	-18%
SVM	FPO	-4%	-6%	-10%
	PSO	-10%	+1%	-0.6%
	GA	-16%	-8.5%	-4%
LR	FPO	-20.5%	-17%	-67%
	PSO	-54%	-57%	-55%
	GA	-54.5%	-57%	-57%
KNN	FPO	+11%	-2%	+3.6%
	PSO	+21%	-4%	+4%
	GA	+10%	-7%	-0.009%

may vary. While FS could improve training efficiency for some datasets, it may require more training cycles for others. Additionally, we have highlighted the limitations of the Flower Pollination Optimization (FPO) algorithm and emphasised the importance of considering multiple evaluation metrics beyond accuracy alone. Finally, we note that this work forms part of our broader research project on healthcare assistant agents, encompassing various aspects, including ethical considerations [9], [10, 11].

References

[1] I. Letteri, AITA: A new framework for trading forward testing with an artificial intelligence engine, in: F. Falchi, F. Giannotti, A. Monreale, C. Boldrini, S. Rinzivillo, S. Colantonio (Eds.), Proceedings of the Italia Intelligenza Artificiale - Thematic Workshops co-located with the 3rd CINI National Lab AIIS Conference on Artificial Intelligence (Ital IA 2023), Pisa, Italy, May 29-30, 2023, volume 3486 of *CEUR Workshop Proceedings*, 2023, pp. 506–511.

[2] I. Letteri, G. D. Penna, G. D. Gasperis, A. Dyoub, Trading strategy validation using forwardtesting with deep neural networks, in: Proceedings of the 5th International Conference on Finance, Economics, Management and IT Business, FEMIB 2023, Prague, Czech Republic, April 23-24, 2023, SCITEPRESS, 2023, pp. 15–25. doi:10.5220/0011715300003494.

[3] I. Letteri, G. D. Penna, G. D. Gasperis, Security in the internet of things: botnet detection in software-defined networks by deep learning techniques, Int. J. High Perform. Comput. Netw. 15 (2019) 170–182. doi:10.1504/IJHPCN.2019.106095.

[4] M. Petwan, K. R. Ku-Mahamud, A review on bio-inspired optimization method for supervised feature selection, International Journal of Advanced Computer Science and Applications 13 (2022). URL: http://dx.doi.org/10.14569/IJACSA.2022.0130516. doi:10.14569/IJACSA.2022.0130516.

[5] I. Letteri, G. D. Penna, L. D. Vita, M. T. Grifa, Mtakdd'19: A dataset for malware traffic detection, in: Proceedings of the Fourth Italian Conference on Cyber Security, Ancona, Italy, February 4th to 7th, 2020, volume 2597 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 153–165. URL: https://ceur-ws.org/Vol-2597/paper-14.pdf.

[6] M. Khushi, K. Shaukat, T. M. Alam, I. A. Hameed, S. Uddin, S. Luo, X. Yang, M. C. Reyes, A comparative performance analysis of data resampling methods on imbalance medical data, IEEE Access 9 (2021) 109960–109975. doi:10.1109/ACCESS.2021.3102399.

[7] M. Sharawi, H. M. Zawbaa, E. Emary, Feature selection approach based on whale optimization algorithm, in: 2017 Ninth international conference on advanced computational intelligence (ICACI), IEEE, 2017, pp. 163–168.

[8] I. Letteri, G. D. Penna, P. Caianiello, Feature selection strategies for HTTP botnet traffic detection, in: 2019 IEEE European Symposium on Security and Privacy Workshops, EuroS&P Workshops 2019, Stockholm, Sweden, June 17-19, 2019, IEEE, 2019, pp. 202–210. doi:10.1109/EUROSPW.2019.00029.

[9] A. Dyoub, S. Costantini, F. A. Lisi, Learning domain ethical principles from interactions with users, Digit. Soc. 1 (2022). doi:10.1007/S44206-022-00026-Y.

[10] A. Dyoub, S. Costantini, I. Letteri, Care robots learning rules of ethical behavior under the supervision of an ethical teacher (short paper), in: P. Bruno, F. Calimeri, F. Cauteruccio, M. Maratea, G. Terracina, M. Vallati (Eds.), Joint Proceedings of the 1st International Workshop on HYbrid Models for Coupling Deductive and Inductive ReAsoning (HYDRA 2022) and the 29th RCRA Workshop on Experimental Evaluation of Algorithms for Solving Problems with Combinatorial Explosion (RCRA 2022) co-located with the 16th International Conference on Logic Programming and Non-monotonic Reasoning (LPNMR 2022), Genova Nervi, Italy, September 5, 2022, volume 3281 of *CEUR Workshop Proceedings*, 2022, pp. 1–8.

[11] A. Dyoub, S. Costantini, F. A. Lisi, I. Letteri, Logic-based machine learning for transparent ethical agents, in: F. Calimeri, S. Perri, E. Zumpano (Eds.), Proceedings of the 35th Italian Conference on Computational Logic - CILC 2020, Rende, Italy, October 13-15, 2020, volume 2710 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020, pp. 169–183. URL: https://ceur-ws.org/Vol-2710/paper11.pdf.