

Random forest classification algorithm for medical industry data

Christodoulos Vlachas¹, Lazaros Damianos¹, Nikolaos Gousetis¹, Ioannis Mouratidis¹, Dimitrios Kelepouris¹, Konstantinos-Filippos Kollias^{1}, Nikolaos Asimopoulos¹ and George F. Fragulis¹*

¹Department of Electrical and Computer Engineering, University of Western Macedonia, 501 00 Kozani, Greece

Abstract. Medical industry produces a significant portion of data whereas by adopting various Machine Learning models it can make accurate predictions about public healthcare that can be generalised. Transfer learning improves traditional machine learning by transferring the knowledge learned in one or more tasks and by using it for learning improvement in a related target task. In the current study, transfer learning with random forests was applied. Four datasets of medical interest obtained from the University of California, Irvine (UCI) Machine Learning Repository were used i.e., the BUPA-Liver Disease Dataset, the Breast Cancer Wisconsin Dataset, the Cleveland Heart Disease Dataset, and the Pima Indians Diabetes dataset. To our knowledge, there has been no study that applied Random Forests and Transfer Learning for these datasets. According to our results, our proposed method could provide significant accuracy rates in terms of diagnosing these disorders. Specifically, the classification accuracy of each dataset was similar or higher compared to the majority of similar studies that applied Random Forests. Limitations and suggestions regarding future research are also presented.

1 Introduction

Machine learning (ML) can be described as a subset of artificial intelligence in which a mathematical model based on “training data” is built to make decisions or predictions without being programmed [1]. It is often called a dynamic prediction tool that will make a difference in disease diagnosis and screening in medical industry. Apart from medical industry, Machine Learning has also been applied in other domains for several purposes such as education applications [29], sign language learning [2] and gaming applications [3]. Machine learning is ideal for examining complex high-dimensional data because of its flexibility and automation compared to conventional approaches. These advantages have led to the development of a wide range of diagnostic algorithms [4]. More specifically, diagnostic algorithms have been applied to cardiac arrhythmias detection from electrocardiograms [5], early and objective autism spectrum disorder (ASD) assessment [6,7], pneumonia diagnosis from chest x-rays [8], and identification of patients who need an intervention according to a knee MRI [9].

A variety of algorithms are offered by Machine Learning whereas the Random Forest algorithm is considered as one of the most representative ones. The Random Forest algorithm has shown exceptional success as a regression and a general-purpose classification method [10]. It has been extensively applied in medical industry for the prediction of tumour T cell antigens [11], in a Bayesian Case velocity model regarding the prediction of COVID-19 in the United

States [12] and for the discovery of an optimal traditional Chinese medicine formula that could contribute to the treatment of Alzheimer’s disease.

Traditional machine learning algorithms such as Random Forests can be improved by Transfer Learning which transfers the knowledge learned in one or more tasks and by using it for learning improvement in a related target task [13].

1.1 Objective

The current study, utilised four datasets of medical interest obtained from the University of California, Irvine (UCI) Machine Learning Repository [14]. Random Forests and Transfer Learning were applied in our study. There have been earlier studies that used these datasets and applied either Random Forests [15–19] or Random Forests combined with other algorithms, as well [20–22]. To our knowledge, there has been no study that applied Random Forests and Transfer Learning for these datasets. Therefore, it can be hypothesised that the application of Random Forests and Transfer Learning could result in a better accuracy concerning the prediction of these disorders.

2 Methodology

2.1 Datasets

The datasets obtained from the University of California, Irvine (UCI) Machine Learning Repository [14] are

* Corresponding author: dece00063@uowm.gr

presented in Table 1. The Liver Disorders Dataset (BUPA) [23] is a UCI Machine Learning Repository data set. It consists of a total of 345 instances and 7 attributes. The data instances consist entirely of males. The first 5 attributes are blood metrics and seem to correlate with liver disorders from excessive daily alcohol consumption. The 6th attribute is the daily consumption of alcohol measured in half-pints (equivalent to the alcoholic beverage consumed). The 7th attribute was intended for use as a classifier for data used as training and testing by the data set researchers. The PIMA India Diabetes Dataset [24] aims at predicting whether certain perceptions are in danger of diabetes, given independent factors. There are a few limitations when choosing these instances from a big database. Specifically, all patients are female of Pima Indian descent who are no less than 21 years of age. Eight variables regarding 768 patients [25] are used to find out if the patient has diabetes. The UCI Heart Disease dataset [26], which consists of 304 instances and 14 attributes aims at predicting heart diseases. Predicting heart diseases is difficult since each person has a different pulse rate, blood pressure, and sugar levels. However, there is a range of values the heart should beat per minute and the level of blood pressure should be measured for every person. Making use of them, heart diseases could be predicted early, proper treatment could be given to the patients with a high chance of complete treatment. The Wisconsin Breast Cancer dataset [27] has 699 instances of 10 attributes each, describing physical sizes from measurements of tumours, either benign or malignant.

Table 1. Datasets.

Dataset name	Instances	Attributes	Details
Liver Disorders Dataset (BUPA)	345	7	- Blood metrics → Liver disorders from excessive daily alcohol consumption - Consists entirely of males
The PIMA India Diabetes Dataset	768	8	- Aims at predicting whether certain perceptions are in danger of diabetes - Consists entirely of females no less than 21 years old
The UCI Heart Disease dataset	304	14	- Predicting heart diseases

The Wisconsin Breast Cancer dataset	699	10	- Describing physical sizes from measurements of tumours, either benign or malignant
-------------------------------------	-----	----	--

2.2 Proposed Approach

Random forests in Python version 3.9 were utilised in this study aiming the classification of each dataset used in this paper. Spyder IDE and the Miniconda Virtual Environment were used. The computer configuration used was a Ryzen 2600x processor with 16GB RAM and a GTX 1660 super GPU. Random forest algorithm was applied for all the values of 5 to 150 estimators (trees) and test size ranged from 0.2 to 0.5 producing the following metrics i.e., Max Accuracy, Training Set (Percentage of the dataset used for training), Trees (Number of trees produced) and Average Accuracy. This process was reiterated a total of five times to verify our results.

All the datasets were processed before random forest implementation and classification process. Initially, instances that contained missing values were removed from the dataset, followed by a conversion of all numeric values to integers.

There was some additional processing regarding the Liver Disorders Data Set (BUPA). Particularly, the seventh attribute of this dataset, which was intended for use as a classifier for data used as training and testing by the data set researchers was removed. Moreover, according to UCI guidelines, the 6th attribute presenting the consumption of alcohol per day was edited. Specifically, for values lower than 3, class 0 was defined, while class 1 was defined for values greater or equal to 3.

The process of building and training the random forest classifier was carried out using Python-based Scikit-learn open-source library [28]. Additionally, NumPy and Pandas libraries were imported for use in numeric operations and dataset management, respectively.

The charts were produced using the library mpl_toolkits, part of the matplotlib library. Three-dimensional charts for each dataset were made depicting the number of trees on the x-axis, the test set on the y-axis, and accuracy on the z-axis.

3 Results and Discussion

Four datasets of medical interest obtained from the University of California, Irvine (UCI) Machine Learning Repository [14] were utilised i.e., the Liver Disorders Data Set (BUPA) [23], the PIMA India Diabetes Dataset [24], the UCI Heart Disease dataset [26] and the Wisconsin Breast Cancer dataset [27]. Random Forests and Transfer Learning were applied.

According to Figure 1 and Table 2 presenting the results of the Liver Disorders Dataset (BUPA), the

highest accuracy (66.65%) was achieved while using 42.5% of the dataset as test data with a relatively small number of trees (60). Earlier studies that applied Random Forests found higher accuracy scores than our study while using different methods for the modification of the original dataset. In [20], an extra column provided by the authors of the BUPA dataset was used as a class selector and a k-fold cross-validation method was applied where the dataset was split into equal subsets. The calculated accuracy was 80%. There was one additional study in which cross-validation and the Stability Selection (SS) method were used [15]. Five sub-datasets were produced from this process and the accuracy varied, ranging from 55% to 84%.

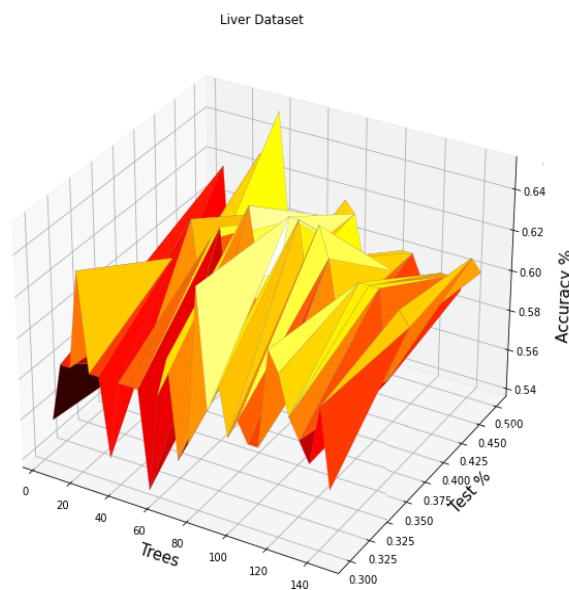


Fig. 1. Liver Disorders Dataset (BUPA) Results.

Regarding the PIMA India Diabetes Dataset, Figure 2 and Table 2 show that the maximum accuracy (80.52%) was achieved while using 30% of the dataset as test data and 90 random trees. When our results are compared with similar studies, it is obvious that a relatively high accuracy was achieved. For example, in [21] RF, C4.5, REPTree, SimpleCart, BFTree, and SVM were applied. Random forest algorithm achieved an accuracy lower than 80.52% with 40 trees while applying a methodology similar to ours. In one additional study [16], an accuracy of 80.08% was achieved with the R tool in Random Forests. These results are similar to ours while a different RF approach was applied.

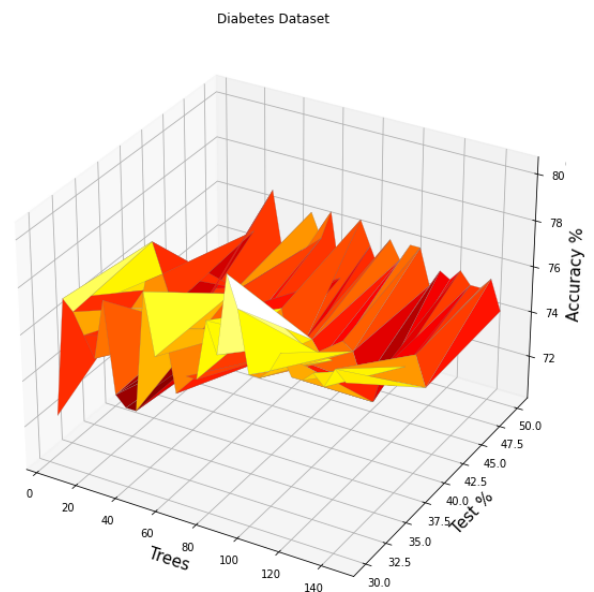


Fig. 2. PIMA India Diabetes Dataset Results.

Concerning the UCI Heart Disease Dataset, we can see in Figure 3 and Table 2 that Random Forest algorithm achieved its highest accuracy (87.91%) while using 30% of the dataset as test data and 75 random trees. When compared to other studies that applied Random Forests, our results are similar (Accuracy = 86.9%) [17] or quite higher than one study that found an accuracy of 82.69% [18] and one study that used Transfer Learning to implement Random Forest Algorithm in WEKA software [22] (Accuracy = 81.51%).

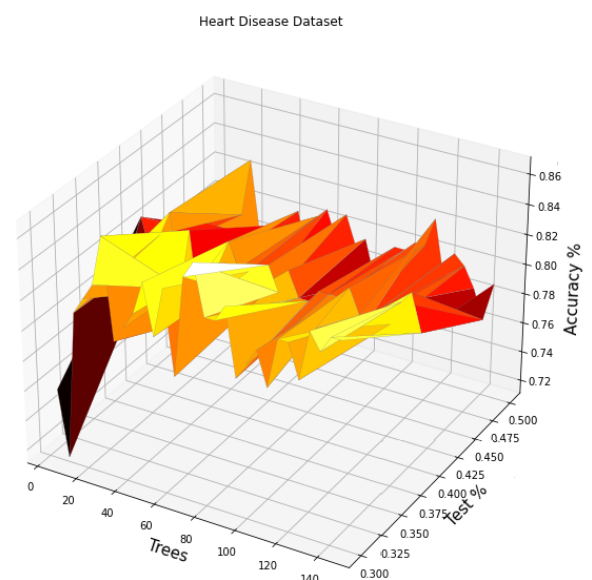


Fig. 3. UCI Heart Disease Dataset Results.

As we can see in Figure 4 and Table 2 depicting the results regarding the Wisconsin Breast Cancer Dataset, the Random Forest algorithm excels in classifying such dataset and achieved peak performance at 96.78% while using 50% of the dataset as test data and 95 random

trees. Using Transfer Learning to implement the Random Forest Algorithm appears to be a significant improvement over other implementations, which present accuracy of 69.23% by applying traditional RF [18]. There was also one study that used Transfer Learning to implement the Random Forest Algorithm in WEKA software [22] finding a classification accuracy that was similar to ours (96.13%). No further improvement can be achieved unless using algorithms with heuristics to optimize the data as presented in other studies [19] which managed to perform a higher accuracy (99.04%).

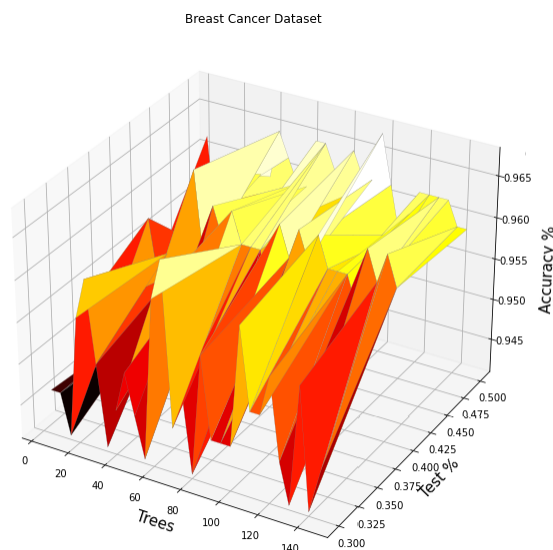


Fig. 4. Wisconsin Breast Cancer Dataset Results.

Table 2. General Result

Dataset name	Max Accuracy	Test Dataset	Number of trees produced	Average Accuracy
Liver Disorders Dataset (BUPA)	66.65%	42.5%	60	65.56%
The PIMA India Diabetes Dataset	80.52%	30%	90	79.22%
The UCI Heart Disease dataset	87.91%	30%	75	86.81%
The Wisconsin Breast Cancer dataset	96.78%	50%	95	96.76%

4 Limitations

During our study, certain limitations were encountered. First, the maximum accuracy was limited, due to the randomness of the trees. We would have a better model if it was fed with values from the previous iteration. By using the Principal Component Analysis (PCA), we could have achieved higher accuracy and a more stable graph representation. This technique increases the interpretability and reduces the dimensionality of the datasets. Another way of increasing our accuracy would be with K-FOLD cross-validation if applied in values of folds ranging from 5 to 10. We could also combine it with the Stability Selection method on the K-FOLD Sub-datasets, to further improve our accuracy. Regarding the classification process efficiency, it could be increased by applying Random Under-Sampling to the original datasets. Lastly, there would be better dataset optimization if algorithms with heuristics were applied.

5 Conclusion

The present study, utilised four datasets of medical interest obtained from the University of California, Irvine (UCI) Machine Learning Repository [14]. Random Forests and Transfer Learning were applied. Our proposed method could provide significant accuracy rates in terms of diagnosing Liver Disorders, Diabetes, Heart Diseases, and Breast Cancer. It also addressed an up-to-date problem of the medical community, since it can predict these disorders accurately, faster than traditional diagnosing methods, and most importantly, with very little computational power and training time.

References

1. Zhang, X.-D. *A Matrix Algebra Approach to Artificial Intelligence*; Springer, 2020;
2. Papatsimouli, M.; Lazaridis, L.; Kollias, K.-F.; Skordas, I.; Fragulis, G.F. Speak with Signs: Active Learning Platform for Greek Sign Language, English Sign Language, and Their Translation. **2020**, doi:10.48550/arXiv.2012.11981.
3. Lazaridis, L.; Papatsimouli, M.; Kollias, K.-F.; Sarigiannidis, P.; Fragulis, G.F. Hitboxes: A Survey About Collision Detection in Video Games. In *Proceedings of the International Conference on Human-Computer Interaction*; Springer, 2021; pp. 314–326.
4. Wilkinson, J.; Arnold, K.F.; Murray, E.J.; van Smeden, M.; Carr, K.; Sippy, R.; de Kamps, M.; Beam, A.; Konigorski, S.; Lippert, C.; et al. Time to Reality Check the Promises of Machine Learning-Powered Precision Medicine. *The Lancet Digital Health* **2020**, 2, e677–e680, doi:10.1016/S2589-7500(20)30200-4.

5. Hannun, A.Y.; Rajpurkar, P.; Haghighpanahi, M.; Tison, G.H.; Bourn, C.; Turakhia, M.P.; Ng, A.Y. Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network. *Nat Med* **2019**, *25*, 65–69, doi:10.1038/s41591-018-0268-3.
6. Kollias, K.-F.; Syriopoulou-Delli, C.K.; Sarigiannidis, P.; Fragulis, G.F. The Contribution of Machine Learning and Eye-Tracking Technology in Autism Spectrum Disorder Research: A Systematic Review. *Electronics* **2021**, *10*, 2982.
7. Kollias, K.-F.; Syriopoulou-Delli, C.K.; Sarigiannidis, P.; Fragulis, G.F. The Contribution of Machine Learning and Eye-Tracking Technology in Autism Spectrum Disorder Research: A Review Study. In Proceedings of the 2021 10th International Conference on Modern Circuits and Systems Technologies (MOCAST); IEEE, 2021; pp. 1–4.
8. Rajpurkar, P.; Irvin, J.; Ball, R.L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.P.; et al. Deep Learning for Chest Radiograph Diagnosis: A Retrospective Comparison of the CheXNeXt Algorithm to Practicing Radiologists. *PLOS Medicine* **2018**, *15*, e1002686, doi:10.1371/journal.pmed.1002686.
9. Bien, N.; Rajpurkar, P.; Ball, R.L.; Irvin, J.; Park, A.; Jones, E.; Bereket, M.; Patel, B.N.; Yeom, K.W.; Shpanskaya, K.; et al. Deep-Learning-Assisted Diagnosis for Knee Magnetic Resonance Imaging: Development and Retrospective Validation of MRNet. *PLOS Medicine* **2018**, *15*, e1002699, doi:10.1371/journal.pmed.1002699.
10. Biau, G.; Scornet, E. A Random Forest Guided Tour. *Test* **2016**, *25*, 197–227.
11. Jiao, S.; Zou, Q.; Guo, H.; Shi, L. ITTCA-RF: A Random Forest Predictor for Tumor T Cell Antigens. *Journal of translational medicine* **2021**, *19*, 1–11.
12. Watson, G.L.; Xiong, D.; Zhang, L.; Zoller, J.A.; Shamshoian, J.; Sundin, P.; Bufford, T.; Rimoin, A.W.; Suchard, M.A.; Ramirez, C.M. Fusing a Bayesian Case Velocity Model with Random Forest for Predicting COVID-19 in the US. *Available at SSRN 3594606* **2020**.
13. Torrey, L.; Shavlik, J. Transfer Learning Available online: <https://www.igi-global.com/chapter/transfer-learning/www.igi-global.com/chapter/transfer-learning/36988> (accessed on 27 December 2021).
14. UCI Machine Learning Repository Available online: <https://archive.ics.uci.edu/ml/index.php> (accessed on 27 December 2021).
15. Akyol, K.; Gültepe, Y. A Study on Liver Disease Diagnosis Based on Assessing the Importance of Attributes. *IJISA* **2017**, *9*, 1–9, doi:10.5815/ijisa.2017.11.01.
16. Subbaiah, S.; Kavitha, M. Random Forest Algorithm for Predicting Chronic Diabetes Disease. *Advancements in Applications of Microbiology and Bioinformatics in Pharmacology* **2020**, *08*, 4–8.
17. Pal, M.; Parija, S. Prediction of Heart Diseases Using Random Forest. *J. Phys.: Conf. Ser.* **2021**, *1817*, 012009, doi:10.1088/1742-6596/1817/1/012009.
18. Boinee, P.; Angelis, R.D.; Foresti, G.L. Meta Random Forests. *World Academy of Science, Engineering and Technology* **2008**, *18*, 1148–1157.
19. Saqib, P.; Qamar, U.; Aslam, A.; Ahmad, A. Hybrid of Filters and Genetic Algorithm - Random Forests Based Wrapper Approach for Feature Selection and Prediction. In Proceedings of the Intelligent Computing; Arai, K., Bhatia, R., Kapoor, S., Eds.; Springer International Publishing: Cham, 2019; pp. 190–199.
20. Haque, Md.R.; Islam, Md.M.; Iqbal, H.; Reza, Md.S.; Hasan, Md.K. Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder. In Proceedings of the 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2); February 2018; pp. 1–5.
21. Benbelkacem, S.; Atmani, B. Random Forests for Diabetes Diagnosis. In Proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS); April 2019; pp. 1–4.
22. Ali, J.; Khan, R.; Ahmad, N.; Maqsood, I. Random Forests and Decision Trees. *IJCSI International Journal of Computer Science Issues* **2012**, *9*, 272–278.
23. UCI Machine Learning Repository: Liver Disorders Data Set Available online: <https://archive.ics.uci.edu/ml/datasets/liver+disorders> (accessed on 27 December 2021).
24. UCI Machine Learning Repository: Diabetes Data Set Available online: <https://archive.ics.uci.edu/ml/datasets/diabetes> (accessed on 27 December 2021).
25. PIMA Indian Diabetes Prediction. Predicting the Onset of Diabetes | by Ishan Choudhary | Towards Data Science Available online: <https://towardsdatascience.com/pima-indian-diabetes-prediction-7573698bd5fe> (accessed on 27 December 2021).
26. UCI Machine Learning Repository: Heart Disease Data Set Available online: <https://archive.ics.uci.edu/ml/datasets/heart+dise> (accessed on 27 December 2021).
27. UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set Available online: [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)) (accessed on 27 December 2021).
28. Sklearn.Ensemble.RandomForestClassifier Available online: <https://scikit-learn/stable/modules/generated/sklearn.ensemble>

- .RandomForestClassifier.html (accessed on 21 January 2022).
29. Fragulis, G. F., Papatsimouli, M., Lazaridis, L., & Skordas, I. A. (2021). An Online Dynamic Examination System (ODES) based on open source software tools. *Software Impacts*, 7, 100046.