

Yeasts Automated Classification with Extremely Randomized Forests

Valentina Franzoni(✉) and Yana Kozak

Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy
valentina.franzoni@dmi.unipg.it, yana.kozak@studenti.unipg.it

Abstract. Several pathogenic yeast species are resistant to pharmaceutical agents and have evolved so quickly over time that often they cannot be stopped with antifungal treatments. In particular, Candida species can cause yeast infections in patients' blood or organs, and Candida outbreaks are particularly troublesome, often found in hospitals and healthcare facilities where people are most susceptible because of their weakened immune systems. Candida can live for weeks on walls and other surfaces, such as walls or furniture, becoming a major hidden danger. New, rapid, and reliable methods of classification and spread prevention are needed. In this preliminary work, we have considered off-the-shelf standard classifiers for yeast classification and applied them to Candida classification. Since performances of classifiers based on logistic regression, SVM, and Extremely Randomized Forests were not particularly remarkable, the possibility to improve their precision has been explored, applying different techniques of feature extraction and feature selection. Eight different Candida strains have been analyzed and systematically tested for classification. The experimental results show that classification with Extremely Random Forests and feature extraction can achieve promising results with the best precision in the automatic classification of candida yeasts, thus being a powerful tool for the prevention and treatment of candida yeasts outbreaks.

Keywords: Candida · Machine learning · Bioinformatics · Medical informatics · Yeast classification · Supervised classification

1 Introduction

The study of the classification of life forms is systematic in biology and pharmacology and is generally equivalent to the development of a taxonomy. For centuries, taxonomy referred to the description of various life forms and their components. Despite the usefulness of such a structure, the interrelation of life

Conceptualization and methodology, V.F., Y.K.; software planning and programming, Y.K.; results validation, V.F.; formal analysis, V.F.; original draft preparation, review and editing, V.F.; data visualization, Y.K.; supervision and funding acquisition, V.F.

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12955, pp. 436–447, 2021.
https://doi.org/10.1007/978-3-030-87007-2_31

forms was poorly understood. Following the rise of Darwinism, systematic biology began to develop and replace what was simply a matter of nomenclature born out of Aristotle's understanding of the significance of animate life forms. Today, however, many biologists have entered the molecular realm seeking to understand biochemical changes at the molecular level, and this approach led to the evolution of the classification of life forms toward an elucidation of the mechanism of evolution, and the function of such organisms. Today, by feeding algorithms with input from classification features, this process can be automated with greater accuracy and speed. In this study, we focus on the *Candida* yeast.

Saccharomyces yeasts of the genus Candida comprise over 150 species. Many of the Candida species are found on the skin and mucous membranes of animals and are the most common cause of fungal infections worldwide. Many species are harmless commensals or endosymbionts of hosts, including humans; however, when mucosal barriers are disrupted or the immune system is compromised, they can invade and cause diseases, known as opportunistic infections. The worst situation is when such infection outbreaks happen in hospitals and healthcare facilities, where patients with an already compromised immune system can have dangerous side effects including mortality [9], even from easily tractable infections [14,15].

Candida classification [3]has been refined and diversified over the years, culminating in the current basic classification into eight main strains: *Candida Albicans* (CA), *Candida dubliniensis* (CD), *Candida glabrata* (CG), *Candida krusei* (CK), *Candida lusitaniae* (CL), *Candida Orthopsilosis* (CO), *Candida parapsilosis* (CP), and *Candida tropicalis* (CT). Candida exhibits such mutability and adaptability that makes it truly dangerous and highly pathogenic; an invasive capability that applies even to synthetic materials, e.g., used in prosthetics, replacement of internal organs, catheters.

Candida Albicans (CA) spreads ubiquitously. *Blastospores* and *mycelia* are found in tissues. To this day there is debate as to which form is invasive, depending on the type of acetosis and fluid pollution of colonized tissues. A further criterion of differentiation is the division into serotypes A, B and C. In Europe, serotype A is the most diffuse and, in general, CA causes the vast majority of candida mycoses, with a recent further shift of the pathogen spectrum. CA has the greatest adherence to the mucosa compared to other candida types, the fungus can actively penetrate the epithelium with the help of exoenzymes, where the host cell undergoes lysis by plasma membrane enzymes, in particular proteolytic and lipolytic enzymes. The oral variation of CA is particularly well known as a temporary side effect of chemotherapy for the treatment of cancer and lymphomas.

Candida Glabrata (CG) (former name: Torulopsis glabrata) rarely occurs in culture with other fungi and after CA is the most frequent pathogen of fungal infections of the urinary tract. It does not form pseudomycelia and has antigenic commonality with CA. The germ is often found as a cause of vaginal mycoses.

Candida Krusei (CK) is a weedy yeast involved in chocolate production. CK is a nosocomial emerging fungal pathogen found primarily in the immune-compromised patients and those with hematologic cancers. It has natural resistance to standard antifungal agents, e.g., fluconazole. CK is isolated from saliva, nails, bronchi, feces, and vagina. It is present in endocarditis and causes diarrhea in children. Infections with this fungus have increased in recent years.

Candida Tropicalis (CT) is less pathogenic than CA, forms hyphae and pseudophae, and requires more oxygen. The pathogen can proliferate on mucous membranes without any particular signs of disease being observed. However, it also appears as a pathogen of endocarditis and fungal sepsis. CT is also found in mixed cultures with CA, where is extremely resistant to treatment.

Candida Parapsilosis (CP) (synonymous with C. parakrusei) is found in onychomycosis and dermatomycosis, rarely on mucous membranes, identified as a pathogen of fungal endocarditis.

Candida Orthopsilosis (CO) is a human fungal pathogen belonging to the CP. Species complex. CO annotated genome harbors three putative agglutinin-like sequence (ALS) genes.

Similar or same-appearing strains of Candida can give rise to wide variation depending on the geographic area, organic environment, and conditions. Various types of Candida derived from a single ancestor are capable of genetically mutating to become different strains depending on the need for parasitism occurring in a particular condition, and the intoxicants and acidosis conditions of the intracellular and extracellular medium or fluid.

This study aims at an automated classification of candida strains as they might be identified by a human professional based on specific characteristics. The utility is evident in the areas of diagnosis, pharmacology, both for prevention and for treatment. Rapid and accurate identification of pathogenic yeast species is thus critical for clinical diagnosis because of the high level of induced mortality and morbidity, even after antifungal therapy. To this end, new rapid, high-throughput, and reliable identification methods are needed.

In the broad scenario of classification algorithms, we proceeded in two phases. In the first phase, we exploited several preliminary tests with different algorithms, evaluating and comparing the advantages and disadvantages of each algorithm. The advantages can include also interpretability characteristics, thus it is not trivial that the algorithm with the apparent best performance would be the best for the specialized application in health sciences. In the second phase, we chose to test random trees and extremely random forests in-depth, having the advantage to provide readable classification processes which can be analyzed for evaluation. The experimental results show that classification with Extremely Random Forests can achieve promising results in the automatic classification of candida yeasts, thus being a powerful tool for the prevention and treatment of candida yeasts outbreaks.

In Sect. 2, the original data set preprocessing and preliminary evaluation of basic classifiers is presented. In Sects. 3 and 4, the setup and choice of data, features, and evaluation metrics are explained. In Sects. 6 and 7, results are

discussed and conclusions are proposed, including limits of the preliminary work to be outperformed in future studies.

2 Original Yeast Data Set

We analyze a collection of eight different strains, belonging to opportunistic species of the genus Candida isolated from two Italian hospitals (i.e., Pisa and Udine city hospitals) and included in the *Cemin Microbial Collection* of the Microbial Genetics and Phylogenesis Laboratory of Cemin, a center of excellence on nanostructured innovative materials for chemical, physical and biomedical applications based in the Italian University of Perugia.

Genomic DNA was extracted as indicated by Cardinali et al. [1,2] using *Fourier-transform infrared spectroscopy* (FTIR) [8,12]. FTIR is a technique based on the vibration of molecular forms (e.g., fatty acids, nucleic acids, proteins) that make up a microorganism in response to the stress received by infrared (IR) light at specific wavelengths. The data set includes approximately 1600 descriptors of Candida strains, with the related class label obtained by manual analysis of FTIR data on Candida samples. In FTIR, IR radiations strike the microorganism sample under study and its molecular components (e.g., fatty acids, nucleic acids, proteins). Depending on the wavelength of radiation from which samples are hit, the sample may respond by vibrating and thus originating visible spectra (i.e., our descriptors) [5]. Not all wavelengths are effective with all macromolecules, so we will not have detectable signals for all wavelengths used. The values recorded for the descriptors are related to absorbances, expressed in pure numbers (i.e., without units of measurement, see Table). The range of wavelengths between which the descriptors vary is 4000–400 cm^{-1}. Computerized classification allows studying in a rapid, high-throughput way the features obtained by the FTIR technique, and classify samples in an automated way. The researcher can then supervise the classification and proceed directly to further analysis, e.g. the yeasts phylogenetics, the process of branching lineages in the evolution of life, which reconstruction is fundamental to systematics and involved with reconstructing evolutionary kinship relationships of taxonomic groups of organisms at a methodical level.

3 Data Preprocessing and Preliminary Experiments

3.1 Data Preprocessing

The Candida Albicans (CA) class has been reduced for balancing aims from the starting data set, and the randomly removed samples have been used to test each of the classifiers in the preliminary experimental phase, to understand and evaluate the performance of basic classifiers.

A few other samples have been excluded because belonging to additional underrepresented strains, i.e. MG, CL, and CU. In the original data set three identical samplings have been also found (i.e. numbers 204, 205, and 206 of the

original data set), and removed. After the preprocessing phase, the data set consists of strains CA, CG, CO, CP, and CT for a total of 201 samples. The cleaned and balanced data set is fed as input to the preliminary experimental phase to evaluate the performance of basic classifiers.

3.2 Preliminary Experiments

Three state-of-the-art classifiers have been considered for our preliminary experiments: *Logistic Regression*, *Support Vector Machines (SVM)* and *Extremely Randomized Trees*. In the version available in the *skearn* package, the classifiers have been run on the data set with all the 1660 features, and operating on an 80%-20% random split between the training set and test set, uniformly balancing the split over the classes.

3.3 Logistic Regression

Logistic Regression (LR) [10] is a machine learning technique allowing to find optimal data points in classification problems. The model, used when the value to be classified is categorical, takes the input features and attempts to predict a numerical value. The regression model tests the specifications that best fit the information in the experimented data set.

3.4 Support Vector Machines (SVM)

Support Vector Machines [13] are a particular machine learning model used in classification and regression, it builds a series of hyperplanes in n-dimensional space which aim to reach a suitable separation between classes, i.e. maximizing the functional margin, the distance from data of other classes, and minimize the classification error. A basic concept in SVM is the kernel, linear separation of classes can be obtained directly or with the so-called 'kernel trick' where non-separable data are mapped, by a problem specific kernel function, in a higher-dimensional space, and hyperplanes in that space can discriminate classes not separable in the original data space.

3.5 Extremely Randomized Forests

Random Forest classification models [11] are characterized by a training phase in which many decision trees are built and splitting features are selected with criteria of bagging and a random component [4]. The classification task is operated by all the forest trees and the output class is decided by votes the majority of the trees. In the extremely randomized forests [7]b the process of randomization is strongly enhanced in splitting a tree node, at attribute and cut-point level, as randomization grows the generated tree structure tends to be independent of the output of the training samples. One of the main advantages of extreme randomization is computational efficiency, while accuracy needs to be kept under control by appropriate tuning of randomization parameters.

3.6 Classifiers Parameters Settings

The setting of parameters for each classifier has been tuned after preliminary trials and maintained unchanged throughout further phases of the experiments.

Logistics Regression Parameters. The parameters chosen for the logistic regression classifier are characterized by the *OvR* multiclass specifier, which uses the One-vs-All (OvA, i.e. One-vs-Rest) scheme. Class weight has been set to *balanced*, where values of y are used to automatically adjust weights inversely proportional to class frequencies in the input data; *lbgfs* is the solver to be used in the optimization problem; and *max_iter* is the maximum number of iterations for the solver to converge, set for our case to 8000.

SVM Parameters. For Support Vector Machines, a linear kernel is in our case the best performant, with a random state set to 0 to ensure reproducibility.

Extremely Random Forests Parameters. In the case of Extremely Random Forests the *estimators* parameter, i.e. the number of trees inside the forest, has been set to 200 (which was the best performing size, after trying the increasing sequence 50, 100, 150, 200, 250), as criterion measuring the quality of the split the *gini impurity* has been chosen, finally, *random state* has been set to an integer, to obtain a deterministic behavior during fitting and to assure the reproducibility of the classifier.

3.7 Preliminary Results

The results of the preliminary experiments of basic classifiers, with the described settings and tested on an 80%–20% split are shown in Table 1, where best results are highlighted in **bold** case and worst in *italic* case.

Preliminary results with basic classifiers and preprocessed data are not particularly satisfactory. While the average precision is moderately acceptable, for both SVM and LR is greater than 0.92, on the other hand, the performance is not distributed uniformly among the Candida classes. CO and CP are, in fact, still misclassified in about 15% of the cases.

4 Features and Data Set Performances

To improve the performances of classifiers, we explore the application of techniques that change the characteristics of the data set by *feature selection* i.e., reducing the number of features and the noise of the data by focusing on the most significant ones, or by *feature extraction* i.e., improving the effectiveness of features in classification, by generating new features which are a function of existing ones.

4.1 Feature Selection Using Randomized Trees

A successful technique for assessing the importance of features for selection is based on Extremely Random Trees. Since in our preliminary experiments we obtained an 80% accuracy on the data set, the use of this technique is, therefore, a good and promising starting point in terms of reusing the information that our classifier returns on features that proved to be important, thus improving the results for this specialized classification task. This technique is built on the idea of estimating the relative relevance of features for each strain, observing how they are used in nodes of a decision tree. More specifically, if the feature is used in a low-depth node, it is considered more important because it contributes to decisions on a larger fraction of samples, i.e. on their branches.

The estimation of the predictive ability of each feature is averaged over several randomized trees. To reduce the variance of this estimate, the technique of *Min Decrease in Impurity* (MDI) is used, which consists of averaging the estimate over several random trees.

To exclude a major bias in the experiments, we need to avoid using the test set to evaluate the importance of the features. Therefore, for each split data set, the Extremely Randomized Forest feature selection technique is first applied to the training set only; the projection of the training data on the selected features is then finally used to train the different classifiers.

4.2 Feature Extraction Using Principal Component Analysis

Feature extraction has been implemented with the Principal Component Analysis (PCA) algorithm. PCA is a dimensional reduction technique grouping and combining features to create new relevant ones. After several experiments, the PCA technique generated 25 features out of the original 1660. This choice is an accommodation amidst a likely smaller number of features, yet powerful enough to guarantee better performance. With 25 features, the models performed better than with 20, 15, 10, while the performance remained identical for 30 and 35 features.

Table 1. Results of preliminary experiments: basic classifiers, complete set of 1600 features

Classifier	Avg Precision	Avg Accuracy	Avg Recall	Best (Class)	Worst (Class)
LR	0.924	0.925	**0.913**	**0.976 (CG)**	0.857 (CO)
SVM	**0.928**	**0.930**	**0.913**	0.970 (CG)	0.855 (CP)
ERF	*0.853*	*0.835*	*0.818*	*0.931 (CT)*	*0.763 (CA)*

LR = Logistic Regression, SVM = Support Vector Machine, ERF = Extremely Randomized Forests, Best/Worst Precision Class.

5 Evaluation Metrics

Because the division into training and testing sets is randomized, we should avoid performance metrics that may be biased by a random factor: this objective can be achieved using different techniques. The *Stratification* technique has been implemented: although randomized, the division is accomplished by a stratification, which ensures that the distribution is always proportional, i.e., it maintains the percentage of samples for each class, which is the same for both the training and testing sets. In each subdivision, 80% of the originally random samples are assigned to the training set and the remaining 20% to the test set.

One hundred random subdivisions were generated for each combination of *[algorithm/data set/feature selection/feature extraction]*. Classifiers were trained and tested, and their performances were averaged over the validation scores [6]. The overall average *recall, precision, accuracy, F1* scores were calculated both globally and for each class individually, to allow a more detailed performance analysis.

6 Experiments and Results

As previously described, systematic experiments have been conducted on each combination of classifier, feature extraction, or feature selection technique, with the classifiers parameter settings ascertained in the preliminary experiment. A total of six classifiers, including the basic versions, have been run:

- Logistic Regression basic
- Logistic Regression with Feature Selection
- Logistic Regression with Feature Extraction
- Support Vector Machine basic
- Support Vector Machine with Feature Selection
- Support Vector Machine with Feature Extraction
- Extremely Randomized Forests basic
- Extremely Randomized Forests with Feature Selection
- Extremely Randomized Forests with Feature Extraction.

A 100 folds random data splits have been created for each classifier and performance metrics described in Sect. 5 have been collected and averaged.

6.1 Results

The performance metrics obtained in experiments are shown in Tables 2, 3 and 4 where the best results are highlighted.

Table 2 illustrates the recall of each algorithm for each class, we notice that recall of CP is generally not very satisfactory for each algorithm, even in the feature selection or feature selection version. CA and CT recall are also problems with the basic version of Extremely Randomized Forests where the feature

Table 2. Average recall per class

Classifier	Average recall				
	CA	CG	CO	CP	CT
LR	0.913	0.972	0.905	0.792	0.981
LR_FS	0.935	0.974	0.926	0.820	**0.992**
LR_FE	0.893	0.954	0.851	0.688	0.981
SVM	0.946	0.967	0.909	0.756	0.990
SVM_FS	0.968	**0.993**	0.917	**0.826**	0.982
SVM_FE	0.951	0.957	0.877	0.700	0.987
ERF	0.777	0.952	0.914	0.612	0.836
ERF_FS	0.815	0.968	0.897	0.618	0.846
ERF_FE	**0.969**	0.944	**0.948**	0.806	0.967

LR = Logistic Regression, SVM = Support Vector Machine, ERF = Extremely Randomized Forests, _FS = plus feature selection, _FE = plus feature extraction.

extraction version dramatically improves the recall over 0.967. The average precision per class is shown in Table 3, all classifiers have some difficulty with CO and CP, the precision of basic classifiers starting below 0.868 and improving with *SVM plus feature selection* and *ERF plus feature extraction*, where the best improvement is reached by the latter with a remarkable 0.996 precision of CO compared to a 0.868 of its basic version. A feature-powered version seems not to have a uniform increasing effect in the per-class performance of logistic regression, in some cases worse, e.g., CO and CP, while in others, CA and CG, reach the best performances.

Table 3. Average precision per class

Classifier	Avg Precision			Avg Recall	Avg Precision
	CA	CG	CO	CP	CT
LR	0.951	0.976	0.867	0.857	0.971
LR_FS	**0.969**	**0.980**	0.891	0.884	0.982
LR_FE	0.911	0.948	0.835	0.793	0.962
SVM	0.953	0.970	0.874	0.855	**0.988**
SVM_FS	0.966	0.977	0.900	0.927	0.985
SVM_FE	0.937	0.972	0.854	0.810	0.987
ERF	0.763	0.875	0.828	0.868	0.931
ERF_FS	0.788	0.888	0.836	0.896	0.931
ERF_FE	0.935	0.952	**0.904**	**0.996**	0.97

LR = Logistic Regression, SVM = Support Vector Machine, ERF = Extremely Randomized Forests, _FS = plus feature selection, _FE = plus feature extraction.

Table 4. Average f1-score per class

Classifier	Avg f1-score				
	CA	CG	CO	CP	CT
LR	0.928	0.972	0.878	0.807	0.974
LR_FS	0.949	0.976	0.901	0.834	−0.986
LR_FE	0.896	0.949	0.834	0.719	0.869
SVM	0.947	0.967	0.886	0.782	**0.988**
SVM_FS	**0.965**	**0.985**	0.904	0.859	0.982
SVM_FE	0.941	0.962	0.856	0.728	0.985
ERF	0.762	0.908	0.861	0.685	0.873
ERF_FS	0.793	0.923	0.858	0.706	0.879
ERF_FE	0.949	0.945	**0.920**	**0.875**	0.966

LR = Logistic Regression, SVM = Support Vector Machine, ERF = Extremely Randomized Forests, _FS = plus feature selection, _FE = plus feature extraction.

Best per-class F1-score, see Tables 2, 3 and 4 is again obtained by *SVM plus feature selection* and *ERF plus feature extraction*, this latter shows the greatest increment and best values for CO and CP, while maintaining good performances, from 0.945 and more, for all other classes CA, CG, CP. Since F1-score takes into account both precision and recall, the performance of *ERF plus feature extraction* seems to prevail on the other classifiers.

The comparison of all classifier average performances compared to avg-precision, avg- recall, and avg-f1-score, without regard of any particular class, is shown in Table 5.

SVM in all the three versions (*basic*, *SVM plus feature extraction*, *SVM plus feature selection*) provide an average general performance greater of 90% in all

Table 5. Comparisons of average metrics per classifier

Classifier	Average accuracy	Average recall	Average precision	Average f1-score
LR	0.925	0.913	0.923	0.912
LR_FS	0.940	0.930	0.941	0.929
LR_FE	0.893	0.874	0.890	0.873
SVM	0.930	0.913	0.928	0.914
SVM_FS	**0.950**	**0.937**	0.951	**0.939**
SVM_FE	0.915	0.894	0.912	0.894
ERF	0.835	0.818	0.853	0.820
ERF_FS	0.847	0.829	0.868	0.832
ERF_FE	0.939	0.927	**0.952**	0.931

LR = Logistic Regression, SVM = Support Vector Machine, ERF = Extremely Randomized Forests, _FS = plus feature selection, _FE = plus feature extraction.

the three metrics, but as noticed above, this is misleading since precision and recall are not distributed uniformly over the classes, e.g., in the *almost perfect* recall of 0.993 for class CG which introduce a bias in its average performance. The general average performance of *ERT plus feature extraction* is also scoring over 90% in all three metrics, but considering the detailed per-class performance, this version of ERT obtains the best absolute f1-score (0.920 and 0.875) and the best absolute precision (0.904 and 0.996) for classes CP and CP, which showed to be problematic by the other classifiers.

7 Conclusions

In this work, we have presented a comparative study aimed at automatic Candida strains classification. Candida is responsible for yeast infections and the availability of fast, reliable, and automatic methods of detection and classification are of great interest and impact in healthcare. Techniques of feature selection and feature extraction have been applied to standard basic classifiers and experimented on a FITR (Fourier-transform infrared spectroscopy) based dataset of Candida strains.

The experimental results show that these techniques dramatically improve the original basic classifier. The overall best performant classifier the Extreme Randomized Forests improved with PCA-based feature extraction, while very close performance is reached by Support Vector Machine with a technique of feature selection based on Extreme Randomized Forests.

As pointed out in Sect. 3, describing the dataset preprocessing phase, some Candida strain samples, (i.e. MG, CL, and CU) have been excluded because underrepresented, these exclusions represent a clear limit, but extending classification experiments to these strains will be the object of future work. A topic worthy of further investigation is certainly the potential of Extremely Randomised Forests, which plays a decisive role in both best performant approaches.

Acknowledgments. Authors thank Prof. Gianluigi Cardinali for providing the data set, and Luca Roscini for explaining the basics of the FTIR technique. This work is partially supported by the Italian Ministry of Research under PRIN Project "PHRAME" Grant n. 20178XXKFY.

References

1. Colabella, C., Corte, L., Roscini, L., Shapaval, V., Kohler, A., et al.: Merging FT-IR and NGS for simultaneous phenotypic and genotypic identification of pathogenic *Candida* species. PLOS ONE *12*(12), e0188104 (2017). https://doi.org/10.1371/journal.pone.0188104
2. Cardinali, G., Bolano, A., Martini, A.: A DNA extraction and purification method for several yeast genera. Ann. Microbiol. **51**(1), 121–130 (2001). http://europepmc.org/abstract/AGR/IND23244635
3. Colabella, C., et al.: Merging FT-IR and NGS for simultaneous phenotypic and genotypic identification of pathogenic Candida species. PLOS ONE **12**(12), e0188104 (2017). https://doi.org/10.1371/journal.pone.0188104

4. Dietterich, T.: An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach. Learn. **40**, 139–157 (2000). https://doi.org/10.1023/A:100760751394
5. Essendoubi, M., Toubas, D., Bouzaggou, M., Pinon, J.M., Manfait, M., Sockalingum, G.D.: Rapid identification of Candida species by FT-IR microspectroscopy. Biochimica et biophysica acta **1724**(3), 239–247 (2005). https://doi.org/10.1016/j.bbagen.2005.04.019
6. Franzoni, V., Milani, A.: A semantic comparison of clustering algorithms for the evaluation of web-based similarity measures. In: Gervasi, O., et al. (eds.) ICCSA 2016. LNCS, vol. 9790, pp. 438–452. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-42092-9_34
7. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1
8. Helm, D., Labischinski, H., Schallehn, G., Naumann, D.: Classification and identification of bacteria by Fourier-transform infrared spectroscopy. J. Gen. Microbiol. **137**(1), 69–79 (1991). https://doi.org/10.1099/00221287-137-1-69
9. Hökenek, U.D., Özcan, F.G., Sevdi, M.S., Erkalp, K., Selcan, A.: Mortality predictors in sepsis: a retrospective study. Turk. J. Intensive Care **19**, 82–89 (2021)
10. Hosmer, D.W., Jr., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, vol. 398. Wiley, Hoboken (2013)
11. Kam, T.: Random decision forests. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995, pp. 278–282 (1995)
12. Naumann, D., Helm, D., Labischinski, H.: Microbiological characterizations by FT-IR spectroscopy. Nature **351**(6321), 81–82 (1991)
13. Poggioni, V., Franzoni, V., Zollo, F.: Can we infer book classification by blurbs? In: Basili, R., Crestani, F., Pennacchiotti, M. (eds.) Proceedings of the 5th Italian Information Retrieval Workshop, CEUR Workshop Proceedings, Roma, Italy, 20–21 January 2014, vol. 1127, pp. 16–19. CEUR-WS.org (2014). http://ceur-ws.org/Vol-1127/paper3.pdf
14. Pramodhini, S., Srirangaraj, S., Easow, J.M.: Candiduria-study of virulence factors and its antifungal susceptibility pattern in tertiary care hospital. J. Lab. Phys., 1–7 (2021). https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0041-1730880#info
15. Supatharawanich, S., et al.: Invasive fungal infection in children with acute leukemia and severe aplastic anemia: IFI in children with acute leukemia and saa. Mediterr. J. Hematol. Infect. Dis. **13**(1), e2021039 (2021)