Studying the Potential of Multi-Target Classification to Characterize Combinations of Classes with Skewed Distribution

Arne Schneck* Sven Kalle*† Rüdiger Pryss‡ Winfried Schlee† Thomas Probst‡§ Berthold Langguth† Michael Landgrebe† Manfred Reichert‡
Myra Spiliopoulou*
*Otto-von-Guericke Univ. Magdeburg, †University Hospital Regensburg, ‡Univ. Ulm, §Georg-August Univ. Göttingen

***Abstract*—The identification of subpopulations with particular characteristics with respect to a disease is important for personalized diagnostics and therapy design. For some diseases, the outcome is described by more than one target variable. An example is tinnitus: the perceived loudness of the phantom signal and the level of distress caused by it are both relevant targets for diagnosis and therapy. In this work, we study the potential of multi-target classification for the identification of those screening variables, which separate best among the different subpopulations of patients, paying particular attention to subpopulations with discordant value combinations of loudness and distress. We analyse the screening data of 1344 tinnitus patients from the University Hospital Regensburg, including questions from 7 questionnaires, and report on the performance of our workflow in target separation and in ranking the questionnaires' variables on their discriminative power.**

***Index Terms*—multi-target classification on skewed data; tinnitus handicap; tinnitus loudness; medical mining**

I. Introduction

During patient screening, physicians use agreed-upon questionnaires and medical tests to capture symptoms and assessments that associate with the outcome. The results of this screening process are then used for diagnostics and personalized therapy design. There are diseases that pose particular challenges to this process, especially those with comorbidities or with unclear physiopathological mechansisms: extensive assessments are needed for diagnostics, and a complex outcome, consisting of more than one target variables, must be considered for therapy design. In this study, we propose a mining workflow for multi-target classification and for the characterization of assessments with respect to their predictive power towards an outcome consisting of multiple targets. We report on our results on screening data of tinnitus patients, studying the combination of tinnitus loudness and handicap as multi-targeted outcome.

Tinnitus is a medical condition characterized by the phantom perception of sound in one or both ears. In [1], Baguley et al. report a prevalence of 10-15%. The recent review of Elgoyen et al. highlights the large patient heterogeneity as one of the major reasons for inconsistent results in studies on tinnitus [2]. An example of heterogeneity concerns the interplay between loudness of the tinnitus signal and handicap caused by tinnitus: in their study [3], Hiller and Goebel show that ”loudness and annoyance are discrepant” in some cases, since there are patients whose everyday life is obscured although loudness is low, while other patients do not feel disturbed although their tinnitus signal is loud. Understanding what characterizes patients with such a *discordant* combination of loudness and handicap is important for personalized therapy design, but also to gain insights in the pathophysiological mechanisms of tinnitus.

Tinnitus loudness refers to the the subjective loudness of the tinnitus perception, as rated by the patient. The Tinnitus Sample Case History Questionnaire (TSCHQ) [4] contains an explicit question on the scale of tinnitus loudness, as well as further questions on the nature of the perceived signal. The assessment of tinnitus handicap is the subject of the Tinnitus Impairment Questionnaire (TBF12) [5] (12 questions), but questionnaires associated with mental well-being are also of relevance. They include the Major Depression Inventory (MDI) [6] and the World Health Organization Quality of Life (WHOQOL) questionnaire [7]. In this study, we use the tinnitus loudness question (Q11) of TSCHQ [4] as one target variable, ”`TLoudness`”, and the aggregate value of the 25 questions in the Tinnitus Handicap Inventory (THI) [8] as second target variable ”`THandicap`”.

Our approach is a mining workflow with several steps. To learn models for the two targets of tinnitus loudness and handicap we use multi-target classification, preceded by a target discretization task and an oversampling task to account for infrequent combinations of loudness and handicap values. To assess the importance of specific questions/assessments for target separation, our approach encompasses a task of variable ranking, whereby we generate several models and count the occurrences of a variable in a model.

The paper is organized as follows. In section II we give a short overview of related research. In section III we describe the data used in our analysis. In section IV we present our approach and report on the experimental results in section V. We close the paper with a summary and some open questions in section VI.

II. State of the Art

The screening protocols for the diagnosis of a disease encompass questionnaires and clinical examinations. There are several questionnaires for tinnitus diagnostics, including the Tinnitus Sample Case History Questionnaire (TSCHQ) [4], the Tinnitus Handicap Inventory (THI) [8] and the Tinnitus Questionnaire (TQ) [9]. Questionnaires differ in purpose: for example, THI focusses on assessing the handicap caused by tinnitus, while TSCHQ records anamnesis, medication and loudness as well. There is overlap among the questionnaires, but also inside a questionnaire. For example, more than one of the TQ questions addresses the effects of tinnitus on sleep. Agreement or disagreement among answers to strongly correlated questions can shed insights on how a patient experiences the disease. Hence, we do not skip questions before learning, but identify questions that contribute to class separation in a learned model.

Class separation with respect to more than one classification variable is studied by "multi-target" (or multi-output) classification algorithms. Early algorithms include [10]–[12]. Random forests [13] have been shown to be promising for multi-target classification, to the effect that more elaborate algorithms emerged over the years. A recent overview can be found in [14].

The use of ensembles implies that the contribution of each variable to class separation becomes less clear. This is further exacerbated in random forests, since each tree is learned on a different subset of the original feature space. In [13], [15], Leo Breiman already investigates this problem and proposes measures that quantify the *importance* of a variable for class separation. In [16], Louppe et al provide an elaborate quantification of variable importance. They concentrate on Breiman's "Mean Decrease Impurity" (MDI) but their estimations generalize for further impurity measures. The main emphasis of [16] is in providing a reliable estimation of variable importance, i.e. as the number of randomized trees goes towards infinity. Moreover, they provide estimates both for fully grown trees, i.e. after the end of the learning phase, and for trees grown to depth $q \leq p$, where p is a tree's full depth. In our work, we use a much simpler computation of variable importance within a finite set of random trees, without generalization guarantees. To perform this simplification, we restrict model induction so that all randomized trees are learned on the complete feature space instead of learning each tree in a subspace.

Advances on variable ranking also include methods that assess the relevance of variables *before* model learning, aiming to prune non-predictive variables and to identify correlated ones. A recent example can be found in [17]. The a priori exclusion of screening assessments that are known to be overlapping or correlated is not desirable in our example, since we want to identify those among the correlated questions that contribute mostly to separation. Hence, our workflow encompasses the task of variable ranking *after* learning, whereby we perform ranking on the variables learned by a number of multi-target classifiers.

III. Materials

We use a sample of 1344 tinnitus patients from the University Hospital Regensburg and consider exclusively the assessments of the first screening. The screening encompasses answers to several questionnaires, including the Tinnitus Handicap Inventory (THI) [8], which is a 25-items-questionnaire and is the most widely used instrument for measuring the tinnitus-associated handicap in the daily life of the patient. The Tinnitus Questionnaire (TQ) [9] is another questionnaire containing 52 items; it is frequently used for tinnitus research in Germany. The Tinnitus Sample Case History Questionnaire (TSCHQ) [4] is an assessment instrument with 35 questionnaire items to record demographic variables and clinical characteristics. The Tinnitus Impairment Questionnaire (TBF12) [5] is a short questionnaire to measure the tinnitus-related distress; it contains 12 items. The small Tinnitus Severity (TS) questionnaire consists of 6 items, used to measure different aspects of the tinnitus-related distress on a numeric rating scale. The Major Depression Inventory (MDI) [6] is a standard instrument to assess depressive symptoms; it contains 12 items. The World Health Organization Quality of Life (WHOQOL) is an internationally validated questionnaire to measure the quality of life; it encompasses 26 items [7]. In addition to the questionnaires, an Audiological Examination was also performed to assess the hearing ability of the patients with an audiogram.

The outcome is described by two variables. We derive them by discretizing the TSCHQ variable loudnessdescription-text_screen and the THI variable THI_totalscore_screen. The former variable is a measure for the subjective loudness of the tinnitus perception, as rated by the patient. It ranges between 1 and 100 in steps of 5 units. The variable THI_totalscore_screen is the aggregate value of the 25 questions in the Tinnitus Handicap Inventory (THI) questionnaire and ranges between 0 and 100. Values closer to 100 indicate higher loudness, resp. handicap. We split each value range into two bins, the bin "LOW" containing the values in [0,50), the bin "HIGH" containing the higher values in [50,100]. The resulting discrete variables `TLoudness` and `THandicap` are binary. Of the four combinations total, two are discordant, namely low loudness with high handicap (also denoted as L_H+ hereafter) and high loudness with low handicap (denoted as L+H_).

From this dataset we removed all patients, for whom one or both of the target variables had no value. As next filtering step, we projected away following variables: variables with evident logical errors, variables with undiscretized dates, variables with missing values for more than 5% of the patients. Patients with missing values in one of the retained variables were also removed, as final filtering step. The remaining dataset consists of 629 patients described by 97 variables. The distribution of the targets is depicted on Table I.

IV. Our Approach

Our approach for multi-target classification builds upon following model of the learning problem.

TABLE I
DISTRIBUTION OF THE TARGET VARIABLES IN THE PATIENTS SAMPLE

	TLoudness : LOW	TLoudness : HIGH
THandicap : LOW	69	239
THandicap : HIGH	20	301

Let $\mathcal{T} = \{T_1, \ldots, T_m\}$ be the set of targets and let $L_{T_i} = \{C_{i,1}, \ldots, C_{i,l_i}\}$ be the set of class labels for the target T_i. Let $\mathcal{P}_m$ be the set of all combinations of labels from the m targets. Further, let $s = \{s_1, \ldots, s_m\} \in \mathcal{P}_m$ be a combination of labels from the targets, i.e. $s_i \in L_{T_i}$ for each $i = 1 \ldots m$. We define as *learning focus* (or simply *focus*) the set $S \subseteq \mathcal{P}_m$ of label combinations that are of particular interest for the application. For the tinnitus application, S consists of the two discordant combinations of TLoudness and THandicap, namely L_H+ and L+H_. On Table I we see that L_H+ is infrequent (20 patients), while L+H_ is frequent (239 patients).

Our first objective is to build a set of models that separate well both with respect to $\mathcal{P}_m$ and with respect to the focus S. Our second objective is to derive from these models a set of variables with high contribution to the classification of those instances, whose labels are in the focus S.

A. Outline of our mining workflow

Our approach towards the two objectives of classification and identification of predictive variables encompasses following tasks:

1) Bin construction for the target combinations and oversampling
2) Multi-target classification
3) Assessment of model quality
4) Assessment of a variable's importance
5) Construction of "good" models and variable ranking over those models

The first task encompasses partitioning of the training sample into bins, where each bin covers one combination of values of the target variables. Since some bins may be substantially smaller than others (cf. data distribution among the four combinations in Table I), we perform oversampling to derive equisized bins.

In the following, we describe the subsequent tasks of our workflow.

B. Multi-target classification core

For class separation we use random forests (RF), as proposed in [13]. For a training set D over a feature space F, this algorithm induces multiple CART-based decision trees [18], whereby each tree is learned on $|D|$ instances, randomly drawn with replacement from D. RF considers a random choice of variables from F when inducing each tree. In our approach, however, we force RF to consider the whole of F during tree induction, so that all variables in F are considered with equal prior probability during variable ranking.

We consider two RF-based algorithms for multi-target classification. The first one is a scikit-learn implementation [19] of a multi-target classification algorithm on the basis of random trees[1], proposed in [14]. This algorithm, denoted as MTRF hereafter, builds a single classification model for the m target variables, namely an ensemble of random trees. The second algorithm is an RF-based variant on the "Label Powerset" algorithm proposed in [20], denoted as LPRF hereafter. This algorithm learns one target variable, the values of which are the combinations of values of the m target variables in P_m.

C. Quantification of model quality

To ensure that our mining workflow produces models that separate well across all targets, we distinguish between *global quality* and *focus quality* of a model. The global quality of model M is an m-dimensional vector $\mathbf{q}_{global}(M)$, the i^{th} element of which is the accuracy value achieved by M for the i^{th} target. The focus quality is the m-dimensional vector $\mathbf{q}_{focus}(M)$ encompassing the recall values for the combinations in the focus S, i.e. the number of hits for the focus combinations to the number of instances in the focus; the i^{th} element of this vector represents the recall value achieved by M for the i^{th} target. Although we define the quality vectors on the basis of accuracy, resp. recall, any other quality function, e.g. the F-measure, could be used instead.

Since the instances belonging to classes in S may make only a small portion of the population, we consider two user-defined thresholds, τ_{global} and τ_{focus}. A model is "good" if each element of its global quality vector exceeds τ_{global} *and* each element of its focus quality vector is higher than τ_{focus}.

D. Assessing a variable's importance

To assess the importance of a variable for the separation among the classes in P_m and in the focus S, our mining workflow generates a series of models G. Informally, a variable is deemed to be *important*, if it is used by many models in G. Since a model is a consists of trees, a variable that is used to split the root node or another node close to the root has more influence on class separation than a variable that is used in a split close to the leaf nodes. Hence, our scoring function for a variable's importance takes into account the position of a variable in each tree that used this variable for splitting.

Let $s \in S$ be a combination of target variable values from the focus S, let $M \in G$ be a model and $T \in M$ be a tree induced as part of M. We denote as $f(s, T)$ the set of those nodes in T, which contain a split that involves s, i.e. a split that separates the instances belonging to s from those belonging to other target combinations. We identify the variables used in the splits of the nodes in $f(s, T)$ and compute for each of them v its importance for s over all $T \in M$. To do so, we combine two scoring functions, $avF(v, s, G)$ and $avH(v, s, G)$, defined as follows.

[1] scikit-learn 0.17 RF implementation http://scikit-learn.org/stable/modules/ensemble.html#random-forests, accessed on Feb. 10, 2017.

For a variable v, we define its average frequency with respect to s in the set of models G as:

$$avF(v,s,G) = \frac{\sum_{M\in G}\sum_{T\in M}\sum_{x\in f(s,T)} split(x,v)}{\sum_{M\in G}\sum_{T\in M} 1} \tag{1}$$

where $split(x,v)$ acquires the value 1 if node x is split on v and zero otherwise. Larger values are better.

For a variable v, we define the average height (tree layer) in which it appears as:

$$avH(v,s,G) = \frac{\sum_{M\in G}\sum_{T\in M}\sum_{x\in f(s,T)} split(x,v)\cdot l(x,T)}{\sum_{M\in G}\sum_{T\in M}\sum_{x\in f(s,T)} split(x,v)} \tag{2}$$

where $l(x,T)$ refers to the position/layer of the tree, where x is located, divided by the total height of the tree T. The root of the tree is at the layer 1, a leaf at a layer equal to the tree height. The closer the node x is to the root of T, the more important is the variable on which x is split. Hence, smaller values of $avH()$ are better.

On the basis of those two functions, we define the *importance* of a variable v in a set of models G towards a set of focus combinations s as:

$$importance(v,s,G) = avF(v,s,G) - w * avH(v,s,G) \tag{3}$$

where the contribution of frequently used variables is penalized if the location of these variables is close to the leaves of the trees in the models of G. The weight w regulates the influence of $avH()$. In our work, we have set $w = 0.5$.

This function allows us to either extract all variables with higher importance scores than a threshold, or to select the variables with the top-N scores. In section V we choose the second option and return the top $N = 10$ variables for the focus combinations L_H+ und L+H_.

E. Variable ranking for a choice of models

To identify the variables that have the highest contribution to class separation, we stepwise generate a number of models. However, instead of considering models of arbitrary quality, we discard models, the quality of which is below threshold, and continue the model generation until a user-defined number n of good models is reached. They constitute the set of models G input to the importance function.

To create this G for each of the two classification algorithms MTRF and LPRF, we perform a sequence of runs. In each run, we place the instances into two bins, whereby we oversample the minority classes, according to the first step of our workflow. We first use the one bin for learning and the other one for evaluation, and then switch the bins. Hence, each run outputs two models, the quality of which is evaluated against the two quality thresholds. To increase diversity among the runs, we shuffle the instances, i.e. we assign the instances to two bins randomly without replacement. We continue generating pairs of models until the user-defined number of "good" models n is reached.

V. Experiments

A. Experimental design

We evaluate our approach on the sample described in section III, i.e. for two target variables. We set the focus on the combinations of LOW `TLoudness` and HIGH `THandicap`, denoted as L_H+and of HIGH `TLoudness` and LOW `THandicap`, denoted as L+H_.

For LPRF, we set the threshold for global quality τ_{gq} to 0.8 and the quality threshold for the (two) focus combinations of targets τ_{fq} also to 0.8. For MTRF, we set the corresponding values to 0.88, since it turned out, that models created by MTRF predict better and we only want to create the very best possible models. We set the number of models with quality higher than the thresholds to $n = 20$.

B. Results

The performance values over the $n = 20$ models are depicted on Table II. The second column shows the number of models induced, until 20 models with quality above the thresholds were created.

TABLE II
Global quality and focus quality, averaged over the selected 20 models for each algorithm

	Number of models	*Global quality avg (variance)*	*Focus Quality avg (variance)*
MTRF	20 of 142	0.9060 (0.0005)	0.9323 (0.0005)
LPRF	20 of 112	0.8190 (0.0003)	0.8524 (0.0021)

For each of the combinations L_H+ and L+H_ , we sorted the variables used by LPRF on importance and similarly for MTRF . To select and compare these sets, we have set the performance threshold for MTRF higher than for LPRF. On Table III, we depict the top-10 variables for L_H+, which is the least frequent combination in our data (20 patients, cf. Table I). On Table IV, we similarly show the top-10 variables for L+H_, which is rather frequent in our data (239 patients). Variables considered important by both algorithms are represented in row *Both*, while variables found important by only one algorithm are represented in separate rows *MTRF* and *LPRF*. Q_i represents the ith question of the respective questionnaire.

TABLE III
The top-10 important variables for the combination L_H+

		Top-10 important variables for L_H+
Both	8	THI:{Q10, Q12, Q13, Q16, Q17}, TQ:{Q7, Q10, Q15}
MTRF	2	THI:{Q1, Q23}
LPRF	2	THI:Q21, TQ:Q39

C. Discussion

Our workflow shows very good accuracies for MTRF and LPRF. LPRF induced less models than MTRF in order to build the $n = 20$ good models, but this may be attributed to the higher thresholds we used for MTRF. The higher quality of MTRF is

TABLE IV
The top-10 important variables for the combination L+H_

		*Top-10 important variables for L+H_*
Both	7	THI:{Q10, Q12, Q13, Q16, Q17}, TQ:{Q7, Q15}
MT_{RF}	3	THI:{Q1, Q15, Q25}
LP_{RF}	3	THI:{Q7, Q14, Q21}

not completely unexpected, as it learns all targets separately, while $\mathrm{LP}_{\mathrm{RF}}$ learns combinations. This should be evaluated in more detail though, by usage of e.g. confidence intervals on the achieved accuracies. The similarity of both approaches is underlined by the agreement about the variables characterizing each of the focus combinations.

The top-10 important variables for the two focus combinations L_H+ (infrequent) and L+H_ (frequent) come mostly from the Tinnitus Handicap Inventory (THI). This is not surprising, since `THandicap` is derived from the aggregate score of THI. It is more of interest to check *which* questions are among the top-10: they refer to frustration (Q10), pleasures and responsibilities (Q11, resp. Q12), stress in social relations (Q17), rather than to difficulties in hearing people (THI:Q2), anger (Q3) or confusion (Q4).

Despite the correlation of the target `THandicap` with THI, there are four highly discriminative questions from the Tinnitus Questionnaire (TQ) among the top-10 in L_H+ (cf. Table III), though not in L+H_. TQ contains 52 questions, which are formulated as statements. For example, Q15 states that the tinnitus signal is loud most of the times; the patients answer with ”Agree”, ”Disagree” and ”Partially”. The questions overlap: Q7 states that the tinnitus signal is rather faint. Q10 states that the tinnitus sound is unpleasant, while Q39 is on feeling depressed. The occurrence of these questions in Table III indicates that the answers of the patients are very discriminative for L_H+, while the other patients answer these four questions in a way that does not allow to distinguish between L+H_ and the remaining two classes.

As in THI, the TQ questions present in Table III are on feeling annoyed or distressed. Questions on feeling angry, having difficulties to hear others etc, are not adequately discriminative to reach the top-10 positions. This indicates that the patients experience handicap in very different forms, no form being highly prevalent.

There are some constraints in these results. $\mathrm{MT}_{\mathrm{RF}}$ and $\mathrm{LP}_{\mathrm{RF}}$ are conceptually similar algorithms, so the variability of their findings is not large. Moreover, there has been no correction for oversampling: this affects the reliability of the global/focus quality computations, and thus the choice of the n models for subpopulation characterization. Further, the algorithms learned only over 50% of the data, thus the overall model quality may have been lower than possible. Finally, the ranking of the variables has not been tested statistically. Nonetheless, the lists of the top-10 variables are in agreement with expert insight on which questionnaire questions are informative for the combination of tinnitus loudness and handicap.

VI. Conclusion

We presented a mining workflow for multi-target classification and identification of discriminative variables during patient screening, and we have reported our preliminary results on the classification of screening records of tinnitus patients. Our results for two target variables indicate that the approach can build good models and identify discriminative variables that agree with expert insight. Since the screening involves a very large number of questions from semantically overlapping questionnaires, the identification of discriminative questions can help the physicians focus on specific answers for diagnosis and therapy design.

Our first steps of future work are on the alleviation of some of the identified shortcomings, namely correction for oversampling, induction of random trees on subsets of the feature space (subspaces) and usage of the variable ranking estimates of [16], enhancement of the variable ranking mechanism for global quality vs focus quality with appropriate statistical testing, and experiments with more than two target variables.

References

[1] D. Baguley, D. McFerran, and D. Hall, “Tinnitus,” *The Lancet*, vol. 382, no. 9904, pp. 1600–1607, 2013.

[2] A. Elgoyhen, B. Langguth, D. De Ridder, and S. Vanneste, “Tinnitus: perspectives from human neuroimaging,” *Nature Rev Neurosci*, vol. 16, pp. 632–642, Sept. 2015.

[3] W. Hiller and G. Goebel, “When tinnitus loudness and annoyance are discrepant: audiological characteristics and psychological profile,” *Audiology and Neurotology*, vol. 12, no. 6, pp. 391–400, 2007.

[4] B. Langguth, R. Goodey, A. Azevedo, A. Bjorne, A. Cacace, A. Crocetti, L. Del Bo, D. De Ridder, I. Diges, T. Elbert *et al.*, “Consensus for tinnitus patient assessment and treatment outcome measurement: Tinnitus research initiative meeting, regensburg, july 2006,” *Progress in brain research*, vol. 166, pp. 525–536, 2007.

[5] K. V. Greimel, M. Leibetseder, J. Unterrainer, and K. Albegger, “Can tinnitus be measured? methods for assessment of tinnitus-specific disability and presentation of the tinnitus disability questionnaire,” *Hno*, vol. 47, no. 3, p. 196, 1999.

[6] P. Bech, N.-A. Rasmussen, L. R. Olsen, V. Noerholm, and W. Abildgaard, “The sensitivity and specificity of the major depression inventory, using the present state examination as the index of diagnostic validity,” *Journal of affective disorders*, vol. 66, no. 2, pp. 159–164, 2001.

[7] S. M. Skevington, M. Lotfy, and K. A. O’Connell, “The world health organization’s whoqol-bref quality of life assessment: psychometric properties and results of the international field trial. a report from the whoqol group,” *Quality of life Research*, vol. 13, no. 2, pp. 299–310, 2004.

[8] C. Newman, G. Jacobson, and J. Spitzer, “Development of the tinnitus handicap inventory,” *Archives of Otolaryngology–Head & Neck Surgery*, vol. 122, no. 2, pp. 143–148, 1996.

[9] R. S. Hallam, “TQ – manual of the tinnitus questionnaire – revised and updated, 2008,” http://www.richardhallam.co.uk, 2009.

[10] D. Demšar, S. Džeroski, T. Larsen, J. Struyf, J. Axelsen, M. B. Pedersen, and P. H. Krogh, “Using multi-objective classification to model communities of soil microarthropods,” *Ecological Modelling*, vol. 191, no. 1, pp. 131–143, 2006.

[11] J. Struyf and S. Džeroski, “Constraint based induction of multi-objective regression trees,” in *International Workshop on Knowledge Discovery in Inductive Databases*. Springer, 2005, pp. 222–233.

[12] D. Kocev, C. Vens, J. Struyf, and S. Džeroski, “Ensembles of multi-objective decision trees,” in *European Conference on Machine Learning*. Springer, 2007, pp. 624–631.

[13] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] G. Louppe, “Understanding random forests: From theory to practice,” Ph.D. dissertation, University of Liege, Belgium, 10 2014, arXiv:1407.7502.

[15] L. Breiman, “Manual on setting up, using, and understanding random forests v3. 1,” 2002.

[16] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, “Understanding variable importance in forests of randomized trees,” in *Advances in Neural Information Processing Systems*, 2013, pp. 431–439.

[17] Q. Zou, J. Zeng, L. Cao, and R. Ji, “A novel features ranking metric with application to scalable visual and bioinformatics data classification,” *Neurocomputing*, vol. 173, pp. 346–354, 2016.

[18] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] G. Tsoumakas and I. Katakis, “Multi-label classification: An overview,” *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, 2006, the label powerset algorithm is called PT3.