Covered Information Disentanglement: Correcting Permutation Feature Importance in the Presence of Covariates

João Belo Pereira [1 2] **Diogo Bastos** [1 2] **Erik Stroes** [1] **Evgeni Levin** [1 2]

1. Introduction

Learning biological processes remains one of the most relevant tasks in the medical domain. In many cases, determining the key factors in the development of disease has higher priority than the diagnosis itself since it might dictate or guide potential treatments and research directions. One of the current most popular approaches to measuring feature importance is SHAP (Lipovetsky & Conklin, 2001; Štrumbelj & Kononenko, 2014; Lundberg & Lee, 2017), a game-theoretic approach where the features are seen as ”players” and their marginal contributions to all possible feature subset combinations are measured. A recent work by Kumar et al. (Kumar et al., 2020), exposes some mathematical issues with SHAP and concluded that this framework is ill-suited as a general solution to the problem of quantifying feature importance. Local based methods such as LIME (Ribeiro et al., 2016) and its variants (see e.g. (Singh et al., 2016; Ribeiro et al., 2018.; Guidotti et al., 2018; Pereira et al., 2019)) explain predictions on single instances by building weak, yet explainable models on the neighborhood of these instances. While this achieves a higher prediction transparency for each instance, in this work we are mainly concerned with a more holistic view of importance, which may be more appropriate to guide new research directions and unravel disease mechanisms. Tree-based methods are very commonly selected for this purpose because they compute the impurity or Gini importance (Breiman, 2001). The impurity importance is known to be biased in favor of variables with many possible split points, i.e. categorical variables with many categories or continuous variables (Strobl et al., 2007). A generally accepted alternative to computing the Gini importance is that of permutation importance (Fisher et al., 2018), where the performance of a model is compared on the original data and on data where each feature has been sequentially permuted. There is however, the issue of multicollinearity. When features are highly correlated, then permuting one of them is going to have little effect on the model performance since a great deal of the information provided by this feature is ”covered” by its covariates. One option would be to compute correlation between features and permute correlated features together. However, besides the issue of having to choose an arbitrary threshold for considering features to be correlated enough to share their importances, it leaves out the differentiation of their individual contributions to the final prediction. Motivated by the idea that there is an information overlap between different features, we take an information theoretic approach coupled with Markov Random Fields to disentangle the shared information between features and scale their permutation importance accordingly, which we call Covered Information Disentanglement (CID). We demonstrate how *CID* can recover the right importance ranking on a toy dataset, and discuss its efficacy on the Early stage diabetes risk prediction dataset (Islam et al., 2020).

2. Methodology

2.1. Information Theory background

The *entropy* of r.v. X_i is defined as:

$$H(X_i) \equiv -\sum_{x_i} p(x_i) \log p(x_i). \tag{1}$$

Similarly, the *joint entropy* between r.v.s X_i and X_j is defined as:

$$H(X_i, X_j) \equiv -\sum_{x_i}\sum_{x_j} p(x_i, x_j) \log p(x_i, x_j). \tag{2}$$

The mutual information between r.v.s X_i and X_j, is the relative entropy between the joint entropy and the product distribution $p(x_i)p(x_j)$:

$$I(X_i, X_j) \equiv \sum_{x_i}\sum_{x_j} p(x_i, x_j) \log \frac{p(x_i, x_j)}{p(x_i)p(x_j)}. \tag{3}$$

The conditional entropy of an r.v. X_i given X_j is defined as the conditional distribution averaged over the specific values x_j that X_j can take:

$$H(X_i|X_j) \equiv -\sum_{x_j}\sum_{x_i} p(x_i, x_j) \log p(x_i|x_j). \tag{4}$$

[1] Amsterdam University Medical Center, Meibergdreef 9 1105 AZ, Amsterdam, The Netherlands [2] Horaizon, Marshallaan 2 2625 GZ, Delft, The Netherlands. Correspondence to: João Pereira <j.p.belopereira@amsterdamumc.nl>, Evgeni Levin <e.levin@amsterdamumc.nl>.

Using the definitions above, one can derive properties that resemble those of set theory. In fact, Hu Ting (Ting, 2008) established a formal relation between the information measures and their measure theoretic counterparts. An intuitive representation of these relations can be seen in figure 1.

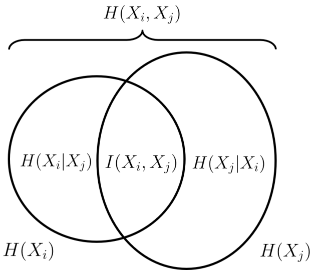

Figure 1. Venn diagram illustrating the relation between different information measure identities.

In order to keep this intuition when generalizing for higher dimensions, one can define the entropy of the ”union” of N features as:

Definition 1. ***Multivariate Union Entropy***

$$H\left(\cup_{i=1}^{N} X_i\right) \equiv -\sum_{x_i} p(x_1,\, ...,\, x_N) log\, p(x_1,\, ...,\, x_N)$$

and using the Inclusion-Exclusion principle, we can define the intersection as:

Definition 2. ***Multivariate Intersection Entropy***

$$H\left(\cap_{i=1}^{N} X_i\right) \equiv \sum_{k=1}^{N} (-1)^{k-1} \sum_{\substack{I\subseteq\{1,\, ...,\, N\};\\ |I|=k}} H(X_{I_1},\, \ldots,\, X_{I_k}).$$

2.2. Permutation Feature Importance

Permutation importance was first introduced by Breiman (Breiman, 2001) in random forests as a way to understand the interaction of variables that is providing the predictive accuracy. It was later expanded by Fisher et al. (Fisher et al., 2018) as a feature importance measure for black box models. Suppose that for a certain feature i in the dataset $\mathbf{X}$ we randomly permute the instances' values, and denote the resultant dataset by $\mathbf{X}_i^\epsilon$. Fisher defines permutation importance, which he refers to as model class reliance, as either the ratio or the difference between the expected loss of $\mathbf{X}_i^\epsilon$ and $\mathbf{X}$:

$$PI_i(f) := \mathcal{L}\left[f\left(\mathbf{X}_i^\epsilon\right)\right] - \mathcal{L}\left[f\left(\mathbf{X}\right)\right] \tag{5}$$

2.3. Covered Information Disentanglement

When there is overlapping information between features, measuring the performance dip when permuting one of the correlated features corresponds to measuring the performance dip by removing the non-mutual information between the feature and its correlates. That is:

$$PI_i\left(f\right) = PI_i^{\mathcal{T}}\left(f\right) - PI_i^{\cup}\left(f\right), \tag{6}$$

where $PI_i^{\mathcal{T}}\left(f\right)$ is the total importance of feature i (the quantity we are interested in) and $PI_i^{\cup}\left(f\right)$ is the performance dip covered by all other variables, which can be computed as:

$$PI_i^{\cup}\left(f\right) = \bigcup_{i\notin\{I\}} \left[PI_i\left(f\right) \cap PI_{\{I\}}\left(f\right)\right]. \tag{7}$$

To compute $PI^{\cup}\left[f\left(\mathbf{X}_i^\epsilon\right)\right]$ would require applying the Inclusion-Exclusion principle and measure the performance dip for all possible feature combinations of size 1 to the number of features. Instead, we note that $PI^{\cup}\left[f\left(\mathbf{X}_i^\epsilon\right)\right]$ intuitively measures the model performance dip when the model is deprived of the information covered by the r.v.s that are correlated with X_i. Motivated by the analogy between set-theory and information measures, we define the fraction of the joint information between a r.v. and the target variable that is ”covered” by the other r.v.s as:

Definition 3. ***Covered information*** *Given a r.v. X_i and a set of distinct r.v.s $\{X_I\}$, $I \subseteq \{1, \ldots N\}\backslash\{i\}$, the Covered information of X_i by X_I w.r.t. Y is defined as:*

$$H_{X_i\cap Y}^{\mathcal{C}(X_I)} = \frac{H\left(X_i \cap Y \cap \{\cup_{j\in I} X_j\}\right)}{H(X_i \cap Y)}$$

Thus, we can use the covered information of X_i by X_I, $I = \{1,\, ...,\, N\}\backslash\{i\}$ w.r.t. Y and re-write equation 6 into:

$$PI^{\mathcal{T}}\left[f\left(\mathbf{X}_i^\epsilon\right)\right] = \frac{PI\left[f\left(\mathbf{X}_i^\epsilon\right)\right]}{1 - H_{X_i}^{\mathcal{C}}\left(X_I\right)}. \tag{8}$$

This means we can approximate the result of permuting all possible combinations of features by computing only the single-feature permutation loss and the covered information of feature X_i by all the others. There is still the issue of computing $H_{X_i\cap Y}^{\mathcal{C}(X_I)}$, since it involves computing $p(X)$. In order to reduce computational time, we will use Markov Random Fields (Koller & Friedman, 2009) yielding the main result of this paper:

Theorem 2.1. *For a Markov Random Field, the covered information of a r.v. X_i by the set of random variables X_I, $I = \{1,\, ...,\, N\}\backslash\{i\}$ w.r.t. Y is given by:*

$$\begin{aligned} &H_{X_i\cap Y}^{\mathcal{C}(X_I)} = \\ &1 + \frac{1}{H(X_i \cap Y)} \mathbf{E}_{\sim p(x_{\sim i,\sim y})}\left[log\left(f\frac{\mathbf{d}^T\mathbf{F}\,\mathbf{e}}{\mathbf{d}^T\mathbf{F}_y\mathbf{F}_{x_i}\mathbf{e}}\right)\right] \end{aligned} \tag{9}$$

where $p(x_{\sim i,\sim y})$ is the joint probability of r.v.s which are neighbors to either X_i or Y, $\mathbf{F}$ is a matrix with the product of joint potential values $\psi_{\mathcal{C}_F}$ for set of cliques

$F : \{X_i, Y \in F\}$; f, $\mathbf{F}_y$ and $\mathbf{F}_{x_i}$ are an entry, column and row of $\mathbf{F}$, respectively, while $\mathbf{d}$ and $\mathbf{e}$ are arrays with the product of potential values $\psi_{\mathcal{C}_D}$, $\psi_{\mathcal{C}_E}$ for set of cliques $D : \{X_i \in D, Y \notin D\}$ and $E : \{X_i \notin E, Y \in E\}$ with fixed X_I.

Proof. You can check the proof in the supplementary material. □

Note how the partition function $\mathbf{Z}$, which is in many cases intractable, is absent in the expectation term. If $H(X_i \cap Y)$ is computed using non-parametric methods, then computing $\mathbf{Z}$ is completely avoidable for the purposes of getting the covered information. In that case, to learn the MRF parameters one can use techniques like score-matching or noise-contrastive estimation (Hyvärinen, 2005; Gutmann & Hyvärinen, 2012).

2.4. Considerations and simplifications

Learning a MRF's network structure is expensive. One popular approach is to use GraphicalLasso which learns the entries of a Gaussian precision matrix by minimizing: $J(\mathbf{\Lambda}) = -log\ det(\mathbf{\Lambda}) + tr(\mathbf{S\Lambda}) + \rho||\mathbf{\Lambda}||_1$, where $\mathbf{\Lambda}$ is the precision matrix, $\mathbf{S}$ is the empirical covariance matrix and ρ acts in analogy to Lasso regularization by penalizing a large number of non-zero precision entries. We can model the potentials using a Gaussian Markov Random Fields whose potentials are $\psi_{s,t}(x_s, x_t) = \exp\left[-\frac{1}{2}\left(x_s\Lambda_{st}x_t + x_s^2\Lambda_{ss} + 2\eta_s x_s\right)\right]$, where $\boldsymbol{\eta} = \mathbf{\Lambda}\boldsymbol{\mu}$ ($\boldsymbol{\mu}$ is the mean vector). However, Gaussian Markov Random fields specify a Markov Random Field over a continuous multivariate distribution and thus the entropy must be replaced by differential entropy, which violates many of the desired properties of discrete entropy. Therefore, we will approximate a continuous distribution with a discrete one $p(x_i) \approx \delta p(\overline{x_i})$, where δ is the bin size and $\overline{x_i}$ is the mean value of the bin, and then carry on with our computations as specified in theorem 2.1. For the case where all bins have the same size, all the δs cancel out.

If we compute the expectation of 9 as the empirical expectation, then the asymptotic complexity becomes $\mathcal{O}(SB^2)$, where S is the number of samples taken for the empirical expectation and B is the maximum between the number of bins used to discretize continuous values and the maximum number of values the discrete features take.

There is some controversy regarding the measure specified by definition 2 as a generalization of mutual information to a number of variables higher than two. The reason is because it may yield negative values. To see this, consider the case of three sets of r.v.s X_i, X_I and Y and suppose there is no correlation between X_i and X_I. If we rewrite the mutual information expression into $I(X_i, Y) - I(X_i, Y|X_I)$, then this expression may become negative when the information provided X_i and Y given a fixed value of X_I is higher than that of $I(X_i, Y)$. This can happen for instance if X_i has no correlation with X_I but knowing X_I introduces a correlation between the two (what is commonly known as "explaining away"). However, in this case there is no overlap between X_i and any other variable and therefore there is no covered information, so the covered information can be set to zero a priori. There are other definitions of generalized mutual information that are always positive such as *total correlation*, *dual total correlation*, *redundancy-synergy index*, *Varadan's synergy* or *partial information decomposition* (Timme et al., 2014), but for our purposes we believe the closed-form simplified expression of 2.1 compensates the potential benefit of using these definitions instead.

3. Experimental Section

To test the *CID* ranking adjustment, we first tested it on a toy dataset where the real importances are known, and then in a real-world medical dataset. For both our tests we used scikit-learn (Pedregosa & et al., 2011) implementation of Extremely Randomized Trees and Graphical Lasso.

3.1. Multivariate Gaussian Test

In order to test if *CID* adjusts the permutation ranking into the correct one, we took 5000 samples from a multivariate Gaussian distribution $X \sim \mathcal{N}(0, \mathbf{\Sigma})$ with 8 features, the last being considered the output variable, and where:

$$\mathbf{\Sigma} = \begin{array}{c} \begin{array}{cccccccc} X_0 & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & Y \end{array} \\ \begin{pmatrix} 1 & 0.4 & 0.1 & 0.1 & 0.1 & 0 & 0 & 0.3 \\ 0.4 & 1 & 0.1 & 0.1 & 0.1 & 0 & 0 & 0.3 \\ 0.1 & 0.1 & 1 & 0.7 & 0.7 & 0 & 0 & 0.4 \\ 0.1 & 0.1 & 0.7 & 1 & 0.7 & 0 & 0 & 0.4 \\ 0.1 & 0.1 & 0.7 & 0.7 & 1 & 0 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0.9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.9 & 1 & 0 \\ 0.3 & 0.3 & 0.4 & 0.4 & 0.4 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Thus, the feature importance is $I(X_2) = I(X_3) = I(X_4) > I(X_1) = I(X_0) > I(X_5) = I(X_6)$. To test the *CID* correction, we performed 20 Shuffle Splits with Extremely Randomized Trees and computed the Gini importances for each feature, as well as the permutation feature importance. We then adjusted the feature importances using the *CID* algorithm. You can compare the rankings in figure 2 As can be seen from figure 2, the feature importance underestimated the importance of X_2, X_3 and X_4 because of their high covariance as expected. The *CID* was able to rectify this ranking and ranked the features in the right order. Moreover, notice how the importance retrieved by the Gini importance underestimated the importance of X_2, presumably because X_3 and X_4 offer nearly as good partitions as X_2 due to their similarity, but also the differences in relative importances are very small in magnitude.

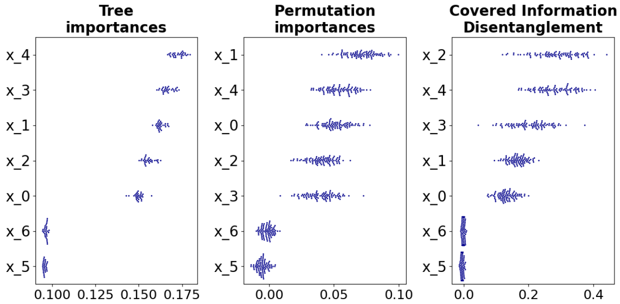

Figure 2. Comparison of the importance ranking on the multivariate gaussian dataset given by from left to right: Tree importance (Gini importance), Permutation Importance, *CID* importance. The ground truth is $I(X_2) = I(X_3) = I(X_4) > I(X_1) = I(X_2) > I(X_5) = I(X_6)$.

3.2. Early stage diabetes risk prediction dataset

In order to test the efficacy of *CID* on a real-world dataset, we used the Early stage diabetes risk prediction dataset (Islam et al., 2020) and applied the same routine as in section 3.1.

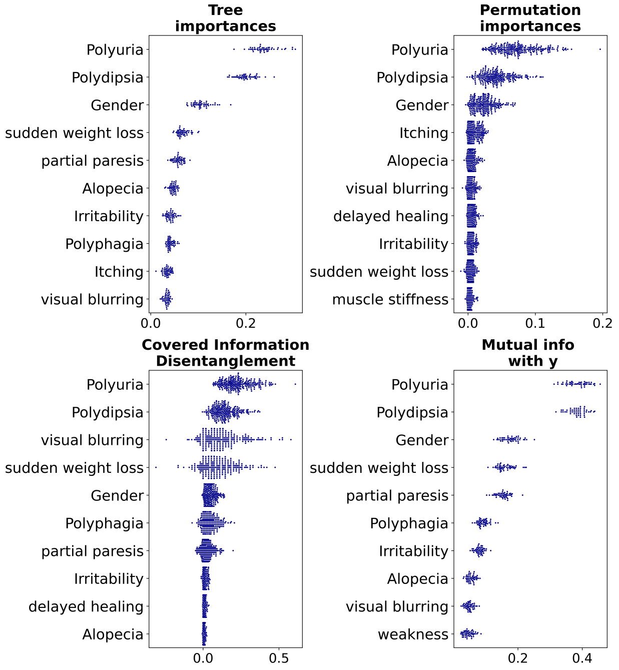

Figure 3. Importance rankings on the Early stage diabetes risk prediction dataset given by Tree importance (Gini importance), Permutation Importance, *CID* importance and Mutual information between the feature and the output.

We also measured the mutual information between each feature and the output to compare the importance rankings with univariate importance measure. A feature that has a high mutual info with the output should be ranked high in importance whereas a feature with low mutual info does not necessarily mean it has low importance because it might have useful information in the multivariate setting. The results for the top 10 importances are depicted in figure 3. Overall, the importances given by *CID* align better with the mutual info than the ones by permutation do. In particular, permutation importance seems to overvalue 'itching', 'Alopecia' and undervalue 'partial paresis', 'Polyphagia' and 'sudden weight loss', while *CID* recovered these into a ranking close to that of the mutual info. The Expert Committee on the Diagnosis and Classification of Diabetes Mellitus mentions that symptoms of marked hyperglycemia include polyuria, polydipsia, weight loss, sometimes with polyphagia, and blurred vision (on the Diagnosis & of Diabetes Mellitus, 2003), all of which *CID* valued highly. *CID* was the only method to attribute high importance to 'visual blurring', a known consequence of unstable blood glucose levels (Jacobsen et al., 2008; Yarbağ et al., 2015). Overall, *CID* ranking seems to align more with the univariate mutual info while still working within a multivariate setting.

4. Discussion and Conclusion

With an increasing reliance on Machine Learning methods to conduct research in impactful domains such as Biology and Medicine, it is more important than ever to achieve model transparency and accurately determine feature relevance. The popular feature permutation method has the advantage of being easy to understand, but its accuracy suffers in the presence of covariates. To address this issue, we suggest a method which uses Information Theory and Markov Random Fields (MRF) to adjust the ranking given by the permutation algorithm and we demonstrate its efficacy on a toy dataset and a real world dataset. These improvements can have a powerful impact in Medical and Biological research since feature importance has widespread applications in these fields for instance to pursue new research directions or the development of drugs. However, inference of an MRF structure is hard and thus we have only explored the case of using Graphical Lasso in conjunction with Gaussian Markov Random Fields. Although this particular implementation is attractive for its scalability and intuitiveness, it might lack sufficient expressive power to model more complex relationships between features and might not be ideally suited for a mix of discrete and continuous data. Besides, since *CID* does not depend on the partition function, it is possible to work with unnormalized distributions whose partition functions are intractable and estimate the parameters using score matching or noise contrastive estimation, an approach not explored here. A further research direction is to use different measures of multivariate mutual information to obtain more accurate values of covered information.

References

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Fisher, A., Rudin, C., and Dominici, F. Model class reliance: Variable importance measures for any machine learning model class, from the "rashomon" perspective. *Computer Science*, 2018.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., and Giannotti, F. Local rule-based explanations of black box decision systems. *http://arxiv.org/abs/1805.10820v1*, 2018.

Gutmann, M. and Hyvärinen, A. Noise- contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

Hyvärinen, A. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Islam, M. M. F., Ferdousi, R., Rahman, S., and Bushra, H. Y. Likelihood prediction of diabetes at early stage using data mining techniques. *Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore,*, 2020.

Jacobsen, N., Jensen, H., Lund-Andersen, H., and Goldschmidt, E. Is poor glycaemic control in diabetic patients a risk factor of myopia? *Acta Ophthalmologica*, 2008.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.

Kumar, E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. A. Problems with shapley-value-based explanations as feature importance measures. *Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria*, 2020.

Lipovetsky, S. and Conklin, M. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pp. 4765–4774, 2017.

on the Diagnosis, T. E. C. and of Diabetes Mellitus, C. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care*, 26(suppl 1):s5–s20, 2003. doi: https://doi.org/10.2337/diacare.26.2007.S5.

Pedregosa, F. and et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Pereira, J., Groen, A. K., Stroes, E. S. G., and Levin, E. Graph space embedding. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pp. 3253–3259, 2019. doi: 10.24963/ijcai.2019/451.

Ribeiro, M., Singh, S., and Guestrin, C. "why should i trust you?" explaining the predictions of any classifier. In *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).*, 2016.

Ribeiro, M. T., Singh, S., and Guestrin., C. Anchors: High-precision model-agnostic explanations. *AAAI*, 2018.

Singh, S., Ribeiro, M. T., and Guestrin, C. Programs as black-box explanations. *http://arxiv.org/abs/1611.07579v1*, 2016.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics*, 8:25, January 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-25.

Štrumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

Timme, N., Alford, W., and et al., B. F. Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *J Comput Neurosci*, 36:119–140, 2014. doi: https://doi.org/10.1007/s10827-013-0458-4.

Ting, H. K. On the amount of information. *Theory of Probability & Its Applications*, 7(4):439–447, 2008.

Yarbağ, A., Yazar, H., Akdoğan, M., Pekgör, A., and Kaleli, S. Refractive errors in patients with newly diagnosed diabetes mellitus. *Pak J Med Sci*, 31(6):1481—–1484, 2015.