EPiC Series in Computing

Volume 70, 2020, Pages 11–20

Proceedings of the 12th International Conference
on Bioinformatics and Computational Biology

Ranking Variable Combinations to Characterize Breast Cancer Subtypes using the IBIF-RF Metric

Isis Narvaez-Bandera and Wandaliz Torres-García

University of Puerto Rico, Mayagüez Campus, PR, 00681 USA
isis.narvaez@upr.edu and wandaliz.torres@upr.edu

Abstract

Gene interactions play a fundamental role in the proneness to cancer. However, detecting and ranking these interactions is a complex problem due to the high dimensionality of genomic data. Hence, we aim to find patterns composed of multiple features to molecularly characterize breast cancer subtypes from the integration of different omics datasets using a data mining approach. To retrieve biological understanding from these computational results, we developed IBIF-RF (Importance Between Interactive Features using Random Forest), a new metric capable of assessing and holistically ranking the importance of genomic interactions without any prior knowledge of key feature combinations. A set of 247 top-performing features from transcriptomic, proteomic, methylation, and clinical data were used to investigate interactive patterns to classify breast cancer subtypes using over 1150 samples. IBIF-RF metric allowed the extraction of 154312, 190481, and 463917 combinations of variables for TCGA, GSE20685, and GSE21653 datasets. Single genes, MLPH and FOXA1, were the most frequently identified variables across all datasets followed by some two-gene interactions such as CEP55-FOXA1 and FOXC1-THSD4. Moreover, IBIF-RF metric allowed the definition of two sets of genes frequently found together (1: FOXA1, MLPH, and SIDT1, and 2: CEP55, ASPM, CENPL, AURKA, ESPL1, TTK, UBE2T, NCAPG, GMPS, NDC80, MYBL2, KIF18B, and EXO1).

1 Introduction

Breast cancer (BC) is a heterogeneous disease and detecting interaction patterns that could lead to new understandings of biological mechanisms is necessary. Detecting pattern interactions is a problem due to the dimensionality of genomic data and the extensive number of possible predictive rules that can be extracted. Nonetheless, there are some techniques used to detect interactions proven to perform well in detecting gene interactions, such as Neural Networks (NNs) [1], Support Vector Machine (SVM) [12], and Random Forests (RF) [6]. Many of these methods have the ability to classify complex classification problems. For example, the NNs method is focused on mimicking the brain's ability to solve problems by connecting a large number of neurons. Though NNs have done well in certain applications, its black-box nature and computational load could make it unattractive for applications with biological data. SVM is another extensively studied classification model that achieves high-performance metrics using hyperplanes and non-probabilistic binary linear classification methodology. This approach has a proven record of working very well in classification problems using complex biological data, however, their outputs can be affected when working with genetic heterogeneity resulting in a lack of interpretability. Moreover, tuning its large amounts of parameters can lead to extensive computational efforts in terms of time and hardware. Among other methods commonly used in the field are the ensemble methods such as RFs. RF models are attractive to study gene interactions since it has several characteristics that fit very well with the requirements of molecular datasets [3]. This type of ensemble can model diverse types of variables with no restrictions on distributional assumptions using a nonlinear approach. Though RF is often categorized as a black-box approach due to its inherent bagging methodology, RF can be interpretable because it can rank the features through the estimation of variable importance measures (VIM) and can evaluate the average marginal effect of a feature in a given class through partial dependency plots (PDPs). Due to these advantages and the capacity of modeling an ensemble of the tree with different random subsets in each, it becomes a solid candidate for discovering interactions.

Therefore, this work uses the ensemble methodology of RF to model breast cancer subtypes (BCs) across thousands of gene expression profiles to focus on measuring those detected interactions and their importance by using a new metric. The assessment of those interactions is critical to interpreting its biological meaning, expand current knowledge, and design further experiments to validate the effects of those significant patterns. It is computationally demanding to calculate all possible rules with its integrative contributions in the model when rules are composed of multiple features each with many possibilities.

In 2014, Deng [4] introduce a framework to extract rules from each tree in the RF ensemble and listing their frequency and associated error which is very useful for interpretation. Nonetheless, this list of rules is often quite extensive and its summarized composition can yield diverse rules requiring long hours of manual inspection to extract biological meaning from them. Moreover, Jones and Linder [7] extended the implementation of PDPs from the average marginal contribution of a feature to estimate the marginal combination of a pair of features including a visualization aid. Visual methods for biological interpretation are of great interest and their method provides an alternative to visualize complex patterns from combinations of features, in their case, a pair of variables. Nonetheless, to be able to implement their algorithm the user must know in advance which pairs of features to analyze. Also, it is hard to interpret partial dependence for high-order interactions. Hence, a new metric called Importance Between Interactive Features using RF (IBIF-RF) capable to assess the most important interactions extracted by RFs was designed and implemented in the context of BCs classification.

2 Methods

The multi-step procedure can be summarized in Figure 1. This approach consists of the following steps: (1) data selection, (2) extraction and integration of important variables, and (3) development and implementation of the IBIF-RF metric.

2.1 Datasets and sources

We studied gene expression, protein, methylation and subtype information from The Cancer Genome Atlas (TCGA) public repository (http://cancergenome.nih.gov/)[9]. The response variable of interest in this study was BCs classification (basal, HER2, luminal A, luminal B, normal) which was provided via PAM50 genomic analysis [10]. For validation purposes, we evaluated the classification performance of extracted variables in the integrative model using

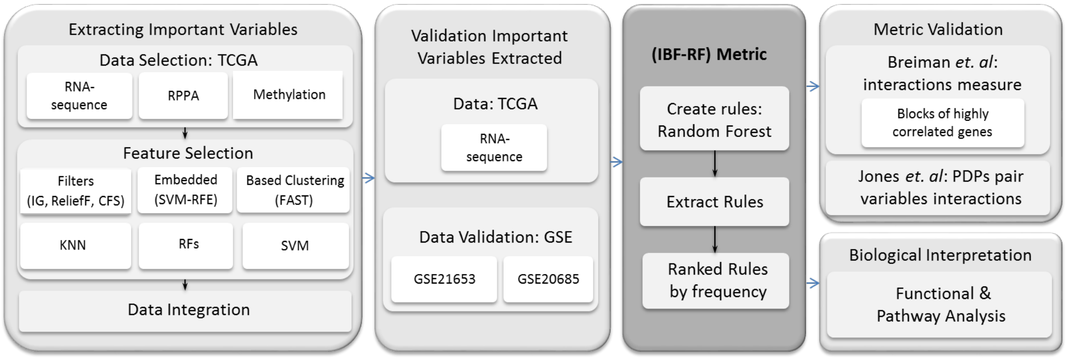

Extracting Important Variables
Data Selection: TCGA
RNA-sequence
RPPA
Methylation
Feature Selection
Filters (IG, ReliefF, CFS)
Embedded (SVM-RFE)
Based Clustering (FAST)
KNN
RFs
SVM
Data Integration
Validation Important Variables Extracted
Data: TCGA
RNA-sequence
Data Validation: GSE
GSE21653
GSE20685
(IBF-RF) Metric
Create rules: Random Forest
Extract Rules
Ranked Rules by frequency
Metric Validation
Breiman et. al: interactions measure
Blocks of highly correlated genes
Jones et. al: PDPs pair variables interactions
Biological Interpretation
Functional & Pathway Analysis

Figure 1: Methodology Framework. The methodology consists of the following steps: extract and integrate important variables, validate important variables, implement the IBIF-RF metric, and interpret the possible biological meaning of results.

two BC microarray datasets from Gene Expression Omnibus repository (GEO), GSE20685, and GSE21653, with 327 and 266 BC samples, respectively.

2.2 Feature Selection

To extract relevant variables that could distinguish between different subtypes, we previously applied five different feature selection (FS) methods: Information Gain (IG), ReliefF, Support Vector Machine based on Recursive Feature Elimination (SVM-RFE), Correlation-based Feature Selection (CFS), and FAST clustering-based using FSelector and OmicsMarkeR R packages and Weka software, and assessed through the following classifiers: k-Nearest Neighbor (KNN), SVM and RF. These proposed methods were selected based on their ability to work with high-dimensional data and their previous use in this field. In this previous work, 247 features from protein, methylation and gene expression were extracted using the TCGA dataset and when integrated using RF, it reached Area Under the Curve (AUC) of 0.86 and an error rate of 0.09 while AUC/error rates values for GSE20685 and GSE21653 were 94.69%/12.23% and 84.92%/15.79%, respectively using only 211 features since not all 247 features were found in these transcriptomics-only platforms [9].

There are many sets of important features that could be evaluated further but due to the computational complexity of extracting rules from RF models, the previously 211 features were considered to extract relevant interactions and evaluate their significant effect predict subtypes. Lastly, we used the reduced list of important features to infer its biological meaning.

2.3 Importance Between Interactive Features using Random Forest (IBIF-RF)

We proposed an IBIF-RF metric that measures the prevalence of a set of features through their recurrence in the ensemble model constructed using RF methodology. The interaction importance will be assessed based on the recurrence of a branch or rule through all the decision trees in the forest towards a particular class x (see Figure 2). To achieve this, an algorithm extracts the rules constructed within the RF classifier, calculates the frequency of the combination of features found across all rules and ranks them as described in Algorithm 1.

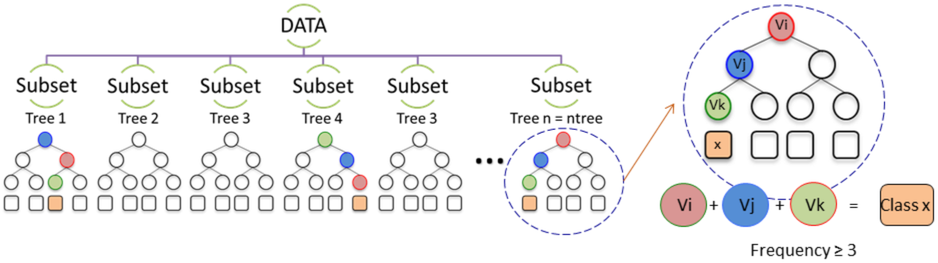


Figure 2: Overview of Importance between Interactive Features using Random Forest (IBIF-RF). The input is the set of important variables extracted from the integration of different types of omics databases. This plot shows an example of the recurrence of a branch (set of features) through all the trees in the forest toward class x. In this example, variables i, j, k are present in three different trees (order does not matter). The metric counts and ranks the prevalence of a variable combination in a rule through their recurrence in the RF model.

The algorithm starts with the construction of an RF model that internally creates subsets of randomized data samples to construct an ensemble of trees. The BootstrapSampling function is an RF internal function that returns a sample that has been taken from N variables with replacement from the full set. The returned samples will be used to build a set of decision trees. The function called BuildsRandomForest runs an RF classifier consisting of a given set of trees, each constructed on a bootstrapped sample set internally using the BootstrapSampling function. These trees are grown and each predictive value is averaged across all trees to provide a final prediction. These trees are translated into classification rules for specific classes. These rules are extracted using the getTree function available in the randomForest R package [8].

Finally, the frequency of these rules composed of features and combinations of features are tallied across all trees to measure its prevalence in the forest. Once all rules were extracted, each rule was split into their individual components (i.e. features and classes) and stored for later use. Duplicated rules within a tree were removed to eliminate redundancies. Finally, this set of ordered variables are stored in a new table, which also indicates the number of the tree and the prediction of the rule. Then, the prevalence of combined variables in a rule is calculated from this table and ordered from highest to lowest frequency to identify interaction patterns (i.e. combination of features) with the highest recurrence. This methodology was implemented using multiple omics datasets available from TCGA and validated using two external gene expression datasets from the GEO data repository (GSE20685 and GSE21653). RF parameters for all models constructed were optimally tuned.

3 Results and Discussion

3.1 Extracted variable combinations

IBIF-RF metric allowed the extraction of 154312, 190481, and 463917 combinations of variables for TCGA, GSE20685, and GSE21653 datasets, respectively using the 211 previously extracted genes. All these combinations of variables were ranked by their frequency in the extracted rules from the entire ensemble. In at least two out of the three datasets in the study, the variable combinations with the highest frequency were those of second and third-order (i.e. 2 and 3 genes in a combination). Although, we obtained rules of higher-order (up to 12 genes) these

```
Input   :
Set $D \leftarrow (X, Y)$;
    $X \leftarrow m$ selected variables $|m = 1, 2, \ldots, M$ ;
    $Y \leftarrow$ response variable levels;
    $T \leftarrow$ Number of trees $|t = 1, 2, \ldots, T$;
STEP   : Build a Random Forest model from dataset $D$ and tuned-up parameters $T$ and $M$;
Require: $RF_t \leftarrow BuildsRandomForest(D_t, T, M)$;
for $t \leftarrow 1$ to $T$ do
    $R(t) \leftarrow$ getTree($RF_t$)
end
Output : $R(t) \leftarrow$ for all decision tree $t$ $|r(rules) = 1, 2, \ldots, R_t$;
         $R_t$ all possible rules in tree $t$;
STEP   : Extract all unique variable combinations in each tree $t$ for all $T$ trees;
for $i \leftarrow 1$ to $T$ do
        for $r \leftarrow 1$ to $R_t$ for a given tree $t$ do
            $CF(r)_(i) \leftarrow$ concatenate features in $(R_t(r))$
        end
        for $x$ in $CF_(t)$ do
            if $x$ is not in $CF_(t)$ then
                $VC(t)$.append($x$)
            end
        end
end
Output : $VC(t)$ set of unique variable combinations for each tree $t$ $|r = 1, 2, \ldots, R_t$;
STEP   : Extract all unique variable combinations in the forest;
Define  : $VCF \leftarrow$ set of unique variable combinations in the forest;
for $i \leftarrow 1$ to $T$ do
    for $y$ in $VC(i)$ do
        if $y$ is not in $VCF$ then
            $VCF$.append($y$)
        end
    end
end
Output : $VCF$;
STEP   : Determine the frequency of unique variable combinations in the forest across all trees;
Define  : $F(VCF) \leftarrow$ frequency of unique variable combinations in the forest;
for $z$ in $VCF$ do
    for $i \leftarrow 1$ to $T$ do
        if $z$ is in $VC(i)$ then
            $F(VCF)[z] = F(VCF)[z] + 1$
        end
    end
end
Output : $F(VCF)$;
STEP   : Rank unique variable combinations in the forest;
Define  : $Rank \leftarrow$ importance ranking of unique variable combinations;
$Rank = VCF$.order(by=$F(VCF)$,descending);
Output : $Rank$;
```

Algorithm 1: Pseudocode IBIF-RF

were less frequent, but still important to enable the prediction of BCs.

We claim that a significant combination of variables extracted from gene expression to discriminate BCs must be found in all three gene expression databases (TCGA, GSE20685, and GSE21653). Consequently, we found 156 variable combinations in common between these datasets. MLPH and FOXA1 rules were found at the top list focusing on discriminating basals from other subtypes as shown in Table 1. These single genes were the most frequently identified variables across all datasets followed by some two-gene interactions. Many of these two-gene combinations included MLPH or FOXA1 with other genes as an interacting pattern with specific expression behaviors characterizing many subtypes (See Table 2).

Table 1: Top common rules extracted by IBIF-RF with their frequency across all datasets.

#	**Rules**	**TCGA**	**GSE 21653**	**GSE 20685**	**Total freq.**
1	MLPH	237	8	56	301
2	FOXA1	109	22	8	139
3	CEP55-FOXA1	3	54	55	112
4	FOXC1-THSD4	11	2	95	108
5	FOXA1-TTK	11	45	46	102
6	MLPH-NOSTRIN	1	2	87	90
7	ASPM-FOXA1	2	47	39	88
8	CENPL-FOXA1	5	11	71	87
9	AURKA-FOXA1	6	43	24	73
10	MLPH-TTK	3	25	45	73
11	ESPL1-FOXA1	12	33	18	63
12	ASPM-FOXC1	3	10	46	59
13	FOXA1-GMPS	10	9	38	57
14	CEP55-MLPH	5	24	25	54
15	FOXC1-KIF18B	10	4	38	52

Table 2: Top genes biological insights using GeneCard (GC) and PubMed (P).

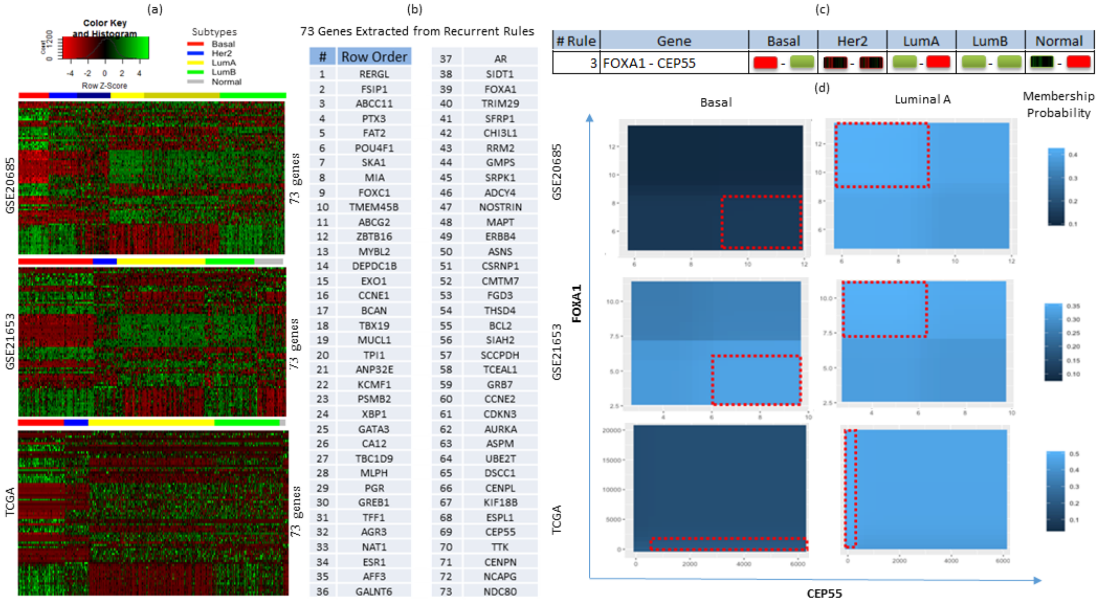
Gene	**GC**			**P**
	Related pathways	**Diseases associated**	**Assoc. w/ BC**	
MLPH	Deregulation of Rab and Rab Effector Genes in Bladder Cancer	Griscelli Syndrome, Type 3 and Osteogenesis Imperfecta, Type Xv	No	3
FOXA1	Embryonic and Induced Pluripotent Stem Cell Differentiation Pathways and Lineage-specific Markers and FOXA1 transcription factor network	Estrogen-Receptor Positive BC and Luminal Breast Carcinoma.	Yes	221
SIDT1	No data available	No data available	No	1
CEP55	Cytoskeletal signaling&DNA damage	No data available	No	7
ASPM	No data available	Microcephaly, Primary, Autosomal Recessive and Autosomal Recessive Primary Microcephaly. Upregulated in several types of cancer: in particular, brain tumors	No, w/ cancer	8
CENPL	Metaphase, Anaphase and Cell Cycle	Seckel Syndrome 1	No	0
AURKA	Integrated BC Pathway and Regulation of PLK1 Activity at G2/M Transition	Colorectal Cancer and Colorectal Adenocarcinoma	No, w/ cancer	177
ESPL1	Metaphase, Anaphase and Cell Cycle	Fallopian Tube Disease&Salpingitis	No	9
TTK	RB in Cancer and DNA Damage	Chronic Polyneuropathy	No, w/ cancer	70
UBE2T	Fanconi anemia pathway and Metabolism of proteins	Complementation Group T and Ube2t-Related Fanconi Anemia.	No	4

3.2 Analysis and discussion of extracted rules

These 156 variable combinations generated in all three datasets are composed of 73 genes. Relevant information about related pathways, associated diseases and supporting literature articles on these 73 common genes were searched (10 of those are listed in Table 2). According to the PubMed engine search on May 16, 2017, we found that 42 out of the 73 genes (57.53%) have more than six published articles related to BC. For instance, the androgen receptor (AR) gene, has over 2000 published papers associated with BC. Also, eight genes (FOXA1, MUCL1, GREB1, TFF1, ESR1, AR, BCL2, GRB7) are directly associated with BC according to GeneCards Human Gene Database (Rebhan et al, 1998). This result supports the sensitivity of our methodology to detect genes that are known to play a key role in BCs identification and understanding. Furthermore, we found genes with little or no publishable track based on this 2017 PubMed search. Nine of those genes (CENPL, RERGL, TBX19, KCMF1, ADCY4, NOSTRIN, CMTM7, SCCPDH, and DSCC1) did not show associated literature in the initial search, whereas 22 of them have between 1 and 5 published studies linked to BC. But since that search in 2017 until now in late 2019, there are some original discoveries linking some of these genes to breast cancer. For example, ADCY4 was detected as a biomarker for breast cancer through epigenetic studies [5] and homolog genes for CENPL such as CENPI have shown correlations with poor prognosis for ER+ cases [11]. Still, to this date, the function of those extracted genes is not well understood. Therefore, these genes are clearly strong candidates for more in-depth explorations of their implications in the BCs.

The expression behavior of the 73 common genes across samples the different five sub-types is shown in Figure 3(a,b). The predictive importance of the 156 common variable combinations from those 73 genes can be observed based on the expression similarity across the two validation datasets shown in Figure 3(a). For example, MLPH and FOXA1 are relevant to differentiate basal subtypes because of their distinct expression levels, which are always less expressed (bright red) for the basals than for any other subtype. Additionally, the third rule, CEP55-FOXA1, characterizes three subtypes: basal, normal and HER2 with distinguishable combinations of expression. From the common variable combinations, we observe two sets of genes with similar behavior (i.e. blocks), and that the combinations of the variables are formed by genes between blocks and not within the same block. The two blocks of genes are formed by FOXA1, MLPH and SIDT1 genes as block 1, and the second by CEP55, ASPM, CENPL, AURKA, ESPL1, TTK, UBE2T, NCAPG, GMPS, NDC80, MYBL2, KIF18B, and EXO1. These genes show evident normalized expression differences across subtypes. The first block is less expressed across all basal samples with a significant change in values when compared with other subtypes while the second group shows less expression in luminal A samples.

Furthermore, to estimate the high expression correlation, within these blocks we calculated Pearson and Spearman correlation metrics for genes within the same blocks. For genes in the first block, MLPH and FOXA1 showed a Pearson correlation value of 0.860, 0.892 and 0.6715 for GSE20685, GSE21653, and TCGA, respectively. Also, the SIDT1 gene with MLPH or FOXA1 shows a strong positive correlation among the GSE datasets (0.74) but lower values for the TCGA (average 0.35). Similarly, for block 2, the correlation between these genes was highest for the GSE datasets (average 0.80) than for TCGA (0.65). These results highlight the high correlation between some genes (i.e. genes blocks) and allow us to suggest that the number of important variables found in the integrative model can be reduced considerably since, between blocks, the genes exhibit similar behavior and may provide the same degree of information regarding the response variable. However, their biological impact must be evaluated as well in terms of the biological relationships among them.



73 Genes Extracted from Recurrent Rules

#	Row Order		
1	RERGL	37	AR
2	FSIP1	38	SIDT1
3	ABCC11	39	FOXA1
4	PTX3	40	TRIM29
5	FAT2	41	SFRP1
6	POU4F1	42	CHI3L1
7	SKA1	43	RRM2
8	MIA	44	GMPS
9	FOXC1	45	SRPK1
10	TMEM45B	46	ADCY4
11	ABCG2	47	NOSTRIN
12	ZBTB16	48	MAPT
13	MYBL2	49	ERBB4
14	DEPDC1B	50	ASNS
15	EXO1	51	CSRNP1
16	CCNE1	52	CMTM7
17	BCAN	53	FGD3
18	TBX19	54	THSD4
19	MUCL1	55	BCL2
20	TPI1	56	SIAH2
21	ANP32E	57	SCCPDH
22	KCMF1	58	TCEAL1
23	PSMB2	59	GRB7
24	XBP1	60	CCNE2
25	GATA3	61	CDKN3
26	CA12	62	AURKA
27	TBC1D9	63	ASPM
28	MLPH	64	UBE2T
29	PGR	65	DSCC1
30	GREB1	66	CENPL
31	TFF1	67	KIF18B
32	AGR3	68	ESPL1
33	NAT1	69	CEP55
34	ESR1	70	TTK
35	AFF3	71	CENPN
36	GALNT6	72	NCAPG
		73	NDC80

# Rule	Gene	Basal	Her2	LumA	LumB	Normal
3	FOXA1 - CEP55	🟥 - 🟩	▨ - ▨	🟩 - 🟥	🟩 - 🟩	▨ - 🟥

Figure 3: (a) Heatmap for Z-scores of gene expression for GSE20685, GSE21653 and TCGA datasets. Rows of each heatmap correspond to the 73 important genes shown in the same order for all datasets. (b) List of the 73 genes extracted from recurrent rules. (c) Recurrent interaction FOXA1 and CEP55 expression levels depiction across subtypes. (d) Marginal interaction plots for FOXA1-CEP55 for Basal and Luminal A subtypes where the ligher the blue shade in the interaction plot the higher the membership probability to belong to that subtype class.

3.3 Validation

For validation purposes, we used two different tools, one offered by Leo Breiman and Adele Cluter [2], and the other by Jones and Linder [7]. First, Breiman and Cutler used an RF method to detect variables interactions. They defined the interaction between two variables as the correlation between them, in the sense that highly correlated variables will have interacting scores. This concept differs from our definition of the interaction of variables, which is the ability of a set of variables to describe a class in a joined manner where we cannot describe a class without one or the other. We applied the code available on the RF web page, for all three databases in the study (TCGA, GSE20685, and GSE21653). The interactions resulting from the Breiman code were, indeed, genes also found within our defined gene blocks, corroborating the strong correlation between them. This is an interesting finding of this work, where highly correlated variables (i.e. genes) can be extracted as important and the RF ensemble model can randomly select any gene and generate predictive rules with different genes, but with the same patterns. The rules extracted through the IBIF-RF metric are a result of a combinatorial process performed by RF to generate the best splits at every node. Due to the highly correlated nature of some genes (grouped by blocks) and the random sampling process of selecting and evaluating variables at each split, we have observed that any gene within the same block (i.e. CEP55 or TTK) can be selected as important with respect to another specific gene (i.e. FOXA1) and still be considered as two different rules.

To further compare the impact of our metric, we used a second tool to validate our results.

Table 3: Prevalence of variable combinations in the RF models.

		TCGA				GSE20685				GSE21653	
#	IR	Var 1	Var 2	#	IR	Var 1	Var 2	#	IR	Var 1	Var 2
1	107	MLPH	FOXA1	1	112	GATA3	ESR1	1	83	GATA3	ESR1
2	99	FOXC1	FOXA1	2	105	THSD4	GATA3	2	79	GATA3	CA12
3	94	ER.alpha	ESR1	3	104	CA12	GATA3	3	76	MYBL2	AURKA
4	90	KRT5	KRT14	4	103	GREB1	ESR1	4	67	MLPH	FOXA1
5	84	CEP55	UBE2T	5	97	ASPM	KIF18B	5	60	MYBL2	ESPL1
6	81	GPR77	ESR1	6	94	ASPM	AURKA	6	55	KIF18B	AURKA
7	79	CDKN3	NDC80	7	91	THSD4	ESR1	7	51	NAT1	GATA3
8	79	KRT17	KRT14	8	89	FOXA1	CAV2	8	50	MYBL2	CENPN
9	78	C6orf97	ESR1	9	82	ASPM	CEP55	9	48	AGR3	GATA3
10	74	DEPDC1B	CEP55	10	78	CA12	ESR1	10	47	ESPL1	AURKA
11	72	EXO1	UBE2T	11	78	GREB1	GATA3	11	46	MYBL2	DSCC1
12	69	MIA	KRT14	12	77	ASPM	ESPL1	12	45	AGR3	ESR1
13	68	EXO1	CEP55	13	77	ASPM	MYBL2	13	45	TBC1D9	GATA3
14	68	AURKA	CEP55	14	74	IGF1R	ESR1	14	45	NCAPG	MYBL2
15	67	AGR3	ESR1	15	71	KIF18B	AURKA	15	44	AR	MLPH
16	66	ASPM	CEP55	16	70	IGF1R	THSD4	16	43	NCAPG	AURKA
17	64	AURKA	UBE2T	17	69	CA12	THSD4	17	43	NAT1	ESR1
18	63	KRT5	MIA	18	67	MLPH	FOXA1	18	41	CA12	ESR1
19	63	MLPH	FOXC1	19	63	NCAPG	ASPM	19	41	TBC1D9	ESR1
20	63	XBP1	FOXA1	20	63	PTX3	CAV2	20	40	TIMELESS	MYBL2

This tool was proposed by Jones and Linder [7] to generate modified PDPs from RF to visualize interactions between pairs of variables. To extract the marginal effect of specific rules we used PDPs and evaluated the behavior of three relevant variable combinations extracted by the IBIF-RF metric: FOXA1-CEP55 (See Figure 3(c)), FOXC1-THSD4, and MLPH-NOSTRIN. We wanted to validate whether the interaction results of these combinations for each subtype were similar to those shown in Table 3. The visualization of FOXA1-CEP55 across all datasets in the study: (a) GSE20685, (b) GSE21653 and (c) TCGA was performed using the plot-pd functions of Edarf R package [7] as shown in Figure 3(d). These plots indicated that: 1) basal occurs when FOXA1 is lowly expressed and CEP55 is highly expressed, 2) luminal A occurs when FOXA1 is highly expressed and CEP55 is lowly expressed and 3) luminal B occurs when both FOXA1 and CEP55 are highly expressed.

These results corroborated the interaction conclusions in this work and lead us to validate that the IBIF-RF metric can rank important interactive patterns, considering all possible rules granting the opportunity to further explore the biological mechanism of these interactions at the experimental level. One big difference between these methods is that the IBIF-RF metric generates common rules for exploration without specifying the specific pair of genes in advance.

4 Concluding Remarks

This work presents the development of the IBIF-RF metric which provides a tool capable to assess interaction importance in a holistic manner without any prior knowledge onto which feature combinations are most important. IBIF-RF metric can rank rules considering all possible ones, this grants the opportunity to pinpoint important interactions and explore the biological meaning of these at the experimental level.

Furthermore, thanks to the evaluation of the IBIF-RF results in a BC case study, two sets of genes that have similar behavior were defined. Also, we were able to infer that several distinct rules are formed by the combination of genes in these sets. The genes forming these two important blocks are FOXA1, MLPH, and SIDT1 genes for the first block, and CEP55, ASPM, CENPL, AURKA, ESPL1, TTK, UBE2T, NCAPG, GMPS, NDC80, MYBL2, KIF18B, and EXO1 for the second one. These results suggest that the number of relevant genes can be reduced noticeably by investigating them by blocks of genes exhibiting similar behavior for a particular response of interest. This reduction could be useful to create clinical panels that are cost-effective. Nonetheless, to improve the current understanding of breast cancer and its subtypes, more in-depth studies of these gene blocks are needed to grasp if the genes as a set have a biological network communication system (i.e. pathway) that can further explain BCs.

On the other hand, seven genes (RERGL, TBX19, KCMF1, NOSTRIN, CMTM7, SCCPDH and DSCC1) extracted still lack reported literature evidence that relates them to BC based on our search in the PubMed database on October 2019. Therefore, it is imperative to further study their biological impact in BC. Lastly, we did not obtain high-frequency values for gene patterns RF rules in comparison with the number of trees in use which can be explained by the random strategy of RF construction. Therefore, the presented metric can be further improved by incorporating normalization step-based on the number of trees used and other important parameters such as the depth of the branches in each tree.

References

[1] Silvio Bicciato, Mario Pandin, Giuseppe Didonè, and Carlo Di Bello. Pattern identification and classification in gene expression data using an autoassociative neural network model. *Biotechnology and bioengineering*, 81(5):594–606, mar 2003.

[2] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[3] Nancy R Cook, Robert Y L Zee, and Paul M Ridker. Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. *Statistics in medicine*, 23(9):1439–53, may 2004.

[4] Houtao Deng. Interpreting tree ensembles with intrees. *arXiv preprint arXiv:1408.5456*, 2014.

[5] Yu Fan, Junhao Mu, Mingquan Huang, Saber Imani, Yu Wang, Sheng Lin, Juan Fan, and Qinglian Wen. Epigenetic identification of adcy4 as a biomarker for breast cancer: an integrated analysis of adenylate cyclases. *Epigenomics*, (0), 2019.

[6] Benjamin A Goldstein, Alan E Hubbard, Adele Cutler, and Lisa F Barcellos. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. *BMC genetics*, 11(1):49, jan 2010.

[7] Zachary Jones and Fridolin Linder. Exploratory data analysis using random forests. In *Prepared for the 73rd annual MPSA conference*, 2015.

[8] A. Liaw and M. Wiener. Classification and Regression by randomForest, 2002.

[9] Isis Narvaez-Bandera, Fernando Sanchez, and Wandaliz Torres-Garcia. Integration of multi omics data for breast cancer subtype classification. In *IIE Annual Conference. Proceedings*, pages 1314–1319. Institute of Industrial and Systems Engineers (IISE), 2017.

[10] Joel S. Parker, Michael Mullins, Maggie C.U. Cheang, Samuel Leung, T David Voduc, Ammi Vickery, Sherri Davies, Christiane Fauron, Zhiyuan Hu Xiaping He, John F. Quackenbush, Inge J. Stijleman, Juan Palazzo, J.S. Marron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, Charles M. Perou, and Philip S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, 27(8):1160–1167, 2009.

[11] Pulari U Thangavelu, Cheng-Yu Lin, Srividya Vaidyanathan, Thu Nguyen, Eloise Dray, and Pascal Duijf. Overexpression of the e2f target gene cenpi promotes chromosome instability and predicts poor prognosis in estrogen receptor-positive breast cancer. *Oncotarget*, 8(37):62167, 2017.

[12] Shen Yuanyuan, Liu Zhe, and J Ott. Detecting gene-gene interactions using support vector machines with L1 penalty. *Bioinformatics and Biomedicine Workshops (BIBMW), 2010 IEEE International Conference on*, pages 309–311, 2010.