Integration of Machine Learning Techniques as Auxiliary Diagnosis of Inherited Metabolic Disorders: Promising Experience with Newborn Screening Data

Bo Lin[1], Jianwei Yin[1(✉)], Qiang Shu[2(✉)], Shuiguang Deng[1], Ying Li[1], Pingping Jiang[2], Rulai Yang[2], and Calton Pu[3]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
{rainbowlin,dengsg,cnliying}@zju.edu.cn, zjuyjw@cs.zju.edu.cn
[2] The Children's Hospital, Zhejiang University School of Medicine, Hangzhou 310058, China
{shuqiang,ppjiang,chsczx}@zju.edu.cn
[3] Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
calton.pu@cc.gatech.edu

Abstract. Tandem mass spectrometry is an advanced biochemical analysis method and has been widely used in screening of inherited metabolic disorders (IMDs). Obtained examination results are filtered by cutoff values and then interpreted based on doctor's knowledge to get diagnoses. However, cutoff-based approaches have difficulties with the correlations of multiple metabolites. Doctor's experiences affect the diagnostic decision-making as well. The rapidly increasing availability of newborn screening data (1.5M cases in this study) enables the application of machine learning (ML) techniques to provide more accurate diagnoses of IMDs compared to simple cutoff values. We investigated two tasks in this study, i.e. complicated patterns between metabolites and better auxiliary diagnostic means. Experimental results show that novel metabolic patterns found in the study are effective and meaningful. Integrating ML techniques with these patterns improved predictive performance compared to existing diagnostic methods, suggesting ML techniques are becoming valuable as auxiliary diagnostic tools.

Keywords: Newborn screening · Machine learning · Metabolic patterns · Auxiliary diagnosis

X. Wang et al. (Eds.): CollaborateCom 2019, LNICST 292, pp. 334–349, 2019.
https://doi.org/10.1007/978-3-030-30146-0_23

1 Introduction

Inherited metabolic disorders (IMDs) are a class of genetic diseases causing mentally disabled, deformity and even death. Systematic screening and treatment to IMDs of newborn can significantly improve prognosis. Research shows that untreated patients tend to spend more money on avoiding neurological sequelae while early intervention is cost-effective over the whole life [18]. Tandem mass spectrometry (MS/MS) is a sensitive, selective and high-throughput technique for concentration detection of various amino acids and acylcarnitine in blood samples, which was first applied to newborn screening in 1990s [15]. A laboratory testing can simultaneously screen out dozens of IMDs, such as amino acid metabolism disorders and fatty acid oxidation disorders, in a few minutes [5]. The existing process of newborn screening is mainly dependent on cutoff-based methods and subjective diagnosis of pediatricians. Setting precise cutoff values for each metabolite or the ratio of two metabolites is the first step to filter out most negative cases. The remaining indistinguishable examination results of MS/MS are interpreted by experienced pediatricians. In practice, cutoff-based methods are hard to deal with complex relationships among metabolites, which bring a large number of false positive cases. As a result, the clinical diagnose still relies on doctor's experience.

In this study, over 1.5M newborn screening data were analyzed by machine learning (ML) techniques, which have proved to be effective for many medical tasks, such as diabetic retinopathy diagnosis [7] and autism spectrum disorder prediction [11]. With enough samples, ML techniques can achieve high performance in the task of disease prediction and act as auxiliary diagnostic means to provide accurate diagnosis. Such diagnostic tool has great social and economic significance. For instance, reducing substantial false alarms not only avoid unnecessary psychological and expenditure burden of families, but improve utilization of medical resources [8]. A refined screening system can be employed in remote districts to enhance the overall quality of medical care in those places. To this end, we aim to answer following two questions: *What is the maximum predictive performance that can be achieved by introducing ML to newborn screening* and *What kind of metabolic patterns can help improve diagnostic accuracy.*

Some related newborn screening projects have explored ML application in IMD diagnosis [2–4]. Compared to these researches, which focus only a few common diseases, we analyzed 16 disorders and evaluated 9 ML algorithms on our dataset. The experimental results demonstrate that more than 20 positive samples are required for a disease to achieve stable performance. Besides common used biomarkers, we discover several metabolites are also contributive to identify diseases. Novel metabolic patterns of their combination outperform existing diagnostic biomarkers. Based on our analysis, ML techniques are effective to be integrated as auxiliary diagnostic tools under certain conditions. Our main contributions are as follows:

- Sixteen IMDs were covered in this study including both common and rare disorders. Extensive experiments with more suitable ML techniques were applied on a large dataset for analysis of practical screening problems.
- We identify a boundary that dividing the applicable situation of ML methods and existing approaches based on the number of positive samples. Possible solutions are provided in both situations.
- We discover novel metabolic patterns in several disorders that achieve higher predictive performance than existing biomarkers.
- Compared to diagnostic methods in existing screening process, we proved integrating ML techniques as auxiliary diagnostic tools can improve predictive accuracy.

To the best of our knowledge, there are few researches that integrating ML techniques into auxiliary diagnosis for dozens of IMDs. We point out strength and weakness of both ML techniques and existing screening approaches in this paper. What's more, further researches can build customized models on the basis of our analyses to improve the screening efficiency (Table 1).

Table 1. List of 16 IMDs investigated in our study.

Abbr.	Disorders
PKU	Phenylketonuria
PTPSD	Tetrahydrobiopterin deficiency
MMA	Methylmalonic acidemia
NICCD	Neonatal intrahepatic cholestasis caused by citrin deficiency
MSUD	Maple syrup urine disease
IVA	Isovaleric acidemia
GA-I	Glutaric acidemia type I
PA	Propionic acidemia
ASS	Citrullinemia type I
VLCAD	Very long-chain acyl-CoA dehydrogenase deficiency
SCAD	Short long-chain acyl-CoA dehydrogenase deficiency
MET	Hypermethioninemia
IBD	Isobutyryl-CoA dehydrogenase deficiency
GA-II	Glutaric acidemia type II
CPT-I	Carnitine palmitoytransferase I deficiency
PRO	Proline acidemia

2 Methods

In this section, we first introduce details of the dataset applied in this study and describe our preprocessing strategies including standardization and train-test split. Various evaluation metrics are then discussed for our imbalanced data.

They are employed reasonably in latter analysis. The experiment was performed on a server with an Intel Xeon E5-2603 1.8 GHz CPU and 16 GB memory. The implementations of ML techniques involved in the experiment are based on the Scikit-learn machine learning framework [16] and imbalanced-learn from its community [14].

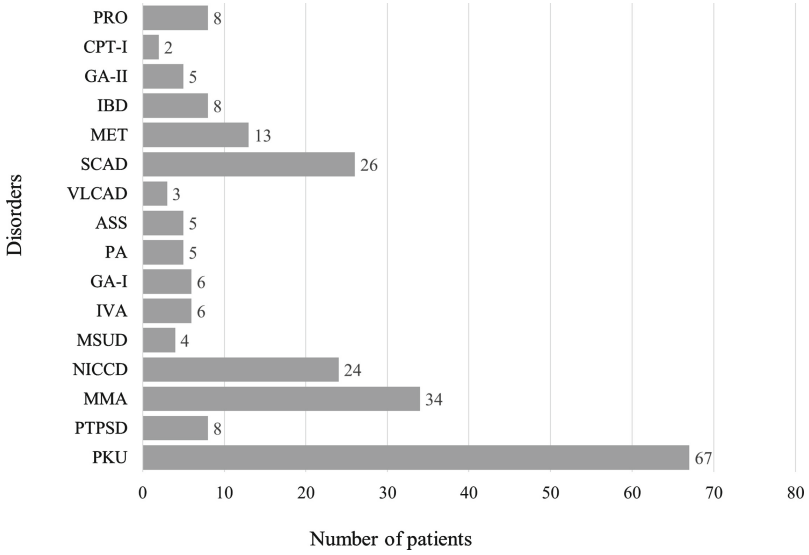

Fig. 1. Number of patients with each disease.

2.1 Data and Preprocessing

The dataset is obtained from the Children's Hospital, Zhejiang University School of Medicine. It consists of 1,506,098 biochemical examination results of neonatus by using MS/MS technique during the period between December, 2011 to December, 2016. Totally 43 biomarkers measured by MS/MS are used in our study (Table A1 in Appendix A), including 11 amino acids, 31 carnitines and a ketone. After excluding diseases with less than two confirmed patients, we obtained 16 IMDs of newborn listed in Tabel 1. Corresponding number of patients in our dataset are shown in Fig. 1. It is worth noting that the actual incidences of disorders are different while total 224 positive cases are only screened by MS/MS. To study the impact of positive sample size on prediction performance, we set two groups, i.e., $G_5 = \{PKU,\ MMA,\ NICCD,\ SCAD,\ MET\}$ and $G_{16} = \{all\ 16\ disorders\}$, whose average number of positive cases are 32.8 and 14, respectively.

One-Versus-Rest (OVR) strategy, a frequently-used multiclass learning approach, was applied to study each disease separately. That is, one disease was selected as the positive class in each time while others were regarded as negative classes. Common sample splitting strategies, e.g., shuffle split and stratified split, may lead insufficient positive samples when setting small training size or bring

redundant negative samples with large size of training set because the splitting almost follows positive-negative sample ratio. Thus, we design an unequal stratified split method called US-split for train-test set generation. The US-split method has a list of class-size parameter $< C, S >$, where S is training size corresponding to target class C. For binary classification problem under OVR strategy, $< C_P, S_P >$ and $< C_N, S_N >$ are two parameter pairs of US-split that controls splitting size of positive and negative classes in training set. In our experiment, we set $S_P^{train} = 0.5$ and $S_N^{train} = 0.05$ indicating sampling half of positive data and 5% of negative data in the partitioning process. These samples were combined as a single training set. Similarly, $S_P^{val} = 0.3$ and $S_N^{val} = 0.1$ were employed for validation set and the remaining were for testing. We ran US-split with replacement for 20 times to generate different train-validation-test splits. All hyperparameters of ML algorithms involved in latter sections were tuned independently in the validation set. Evaluation results of each test set were averaged as the final performance.

A patient who has high metabolite concentration values may strongly affect traditional standardization methods because large values would flat most of samples and make them hard to distinguish. In case of outliers, a robust standardization was applied to training set and test set separately as follows:

$$X_{std} = \frac{X - X_{mid}}{X_{qmax} - X_{qmin}} \tag{1}$$

where X_{mid} is median for current set, X_{qmax} and X_{qmin} are two quantiles that specify the standardizing scale. Due to few outliers, $X_{qmax} = 0.95$ and $X_{qmin} = 0.05$ were chose to cover most of normal samples.

2.2 Evaluation Metrics

We adopted various metrics to evaluate predictive performance from different perspectives. Minimizing false positive rate (FPR) is the primary goal to improve quality of newborn screening while keeping other performances unchanged. Along with this, positive predictive value (PPV), sensitivity (SEN) and specificity (SPE) were considered as basic evaluation metrics. Furthermore, F_β score and *G-mean* were used for comprehensive ability evaluation.

$$F_\beta = (1 + \beta^2)\frac{PPV \times SEN}{\beta^2 \times PPV + SEN} \tag{2}$$

$$G\text{-}mean = \sqrt{SEN \times SPE} \tag{3}$$

Newborn screening dataset is a typical imbalanced dataset and the details will be discussed in Sect. 3. To better evaluating performance of predictors on this kind of dataset, many researches use F_β to avoid being deceived by a single basic metrics with good results due to preference of models. However, small cardinality of patients would bring impact to PPV if the number of false alarms change slightly. In Eq. (2), it is difficult to choose an appropriate β that balancing

PPV and SEN to fairly reflect the model ability. As an alternative metrics that is useful for imbalanced data, *G-mean* in Eq. (3) utilizes SPE to reduce instability brought by PPV. Specificity is the complementary set of FPR so it satisfies our top priority as well. Therefore, *G-mean* is more accord with this study and chose as major evaluation metrics. We also calculate F_1 score as a reference where $\beta = 1$.

3 Results

3.1 Model Comparison

To validate feasibility of introducing ML techniques to newborn screening, we selected nine suitable classification algorithms as shown in Table 2 to compare their performances based on different evaluation metrics. Basically, these algorithms can be categorized into various types such as weighted-bagging and boosting of ensemble methods, linear or nonlinear mapping, tree-based models and so on. Some special configurations of algorithms should be mentioned to avoid misunderstanding. The gradient boosting in our experiment indicates gradient boosting decision tree that using decision trees as weak learners. As for adaptive boosting (adaboost for short), we choose decision stump, or called one-level decision tree [12] as the weak learner.

Table 2. List of nine machine learning algorithms evaluated in our study.

Abbr.	Models
LR	Logistic regression
LDA	Linear discriminant analysis
DT	Decision trees
RF	Random forest
ET	Extremely-randomized tree
GB	Gradient boosting
ADA	Adaptive boosting
SVM	Support vector machine
kNN	k Nearest neighbors

Hyperparameter optimization applied grid search measured by *G-mean* to choose best configuration P_{best} for diverse diseases and models. Other evaluation metrics including PPV, SEN, and SPE were calculated based on the same P_{best}. Interestingly, we set three kernels, i.e., linear, polynomial and radial basis function for SVM classifier in grid search and P_{best} of SVM for different diseases are all configured by the linear kernel. Thus, LSVM is treated as a synonym of the SVM model equipped with linear kernel in the remainder of this paper. During the test, all classifiers are equipped with corresponding P_{best}. Average evaluation results of diseases on group G_5 and G_{16} are treated as their performances.

Table 3. Average evaluation results of nine machine learning algorithms on group G_5 and G_{16}.

Metrics	ADA	LR	GB	DT	ET	LDA	RF	SVM	KNN
GM	**.7201**	**.6114**	**.5519**	.5347	.4988	.4549	.4050	.3235	.2282
	.4549	**.5397**	.3872	.3768	.2973	**.5071**	.1892	.4387	.1225
F1	**.5710**	**.3819**	.3205	**.3818**	.2200	.2730	.3417	.1784	.1769
	.3731	.2768	.2645	.2818	.1543	**.3408**	.1604	**.2976**	.1015
SEN	.5906	**.6585**	.4498	.3937	.3523	**.8851**	.2679	**.9130**	.1146
	.3814	**.7066**	.3301	.3053	.2218	**.7399**	.1327	**.7647**	.0712
SPE	**.9994**	.8563	.9460	.9980	.9688	.5277	**.9999**	.3636	**.9999**
	.9998	.7714	.9763	.9938	.9806	.7282	**.9999**	.6297	**.9999**
PPV	**.7013**	.4877	.3715	**.4905**	.2791	.2747	**.6083**	.1929	.4658
	.4408	.3031	.2962	**.3299**	.2039	**.3420**	.2614	.3016	.2164

Note: The Metrics GM, F_1, SEN, SPE, PPV are G-mean, F_1 score, sensitivity, specificity, positive predictive value, respectively

As we can see in Table 3, the top three best performances for each metrics are in bold. The result in the upper line represents the performance on G_5 and the lower line is on G_{16}. There are some conclusions can be inferred according to the observation of experimental results.

(a) In ensemble-based methods, boosting models especially adaboost have good generalization on G_5, but bagging methods including extremely-randomized tree and random forest perform relatively poor. The reason is that positive cases are insufficient to represent the real distribution. Inconsistent probability distributions are estimated by base classifiers, which affect bagging methods. Their diverse opinions tend to misclassify samples hovering over the border because of the independence between these base classifiers. That is why sensitivity becomes a weakness for bagging models. As for boosting methods, although their base estimators have high bias, the misclassification is considered and passed to next iteration to repair errors. This propagation pays more attention to indistinguishable samples to avoid missed diagnosis. Some false alarms occur but with limited sensitivity degradation, hence keeping high comprehensive performance.

(b) Unlike ensemble methods, linear models such as LR, LDA and LSVM behave well in G_{16}. Specifically, linear models have the three best sensitivity yet their specificity are the worst. The reason is these models have a less complex decision boundary. Under the constraint of avoiding missed diagnosis, they naturally bring false alarms when searching more positive cases. Also, insufficient positive samples reduce generalization of trained models. No matter how to partition training and testing set, the diversity of positive cases is still low. Theoretically, the cutoff-based method is a kind of naïve linear model. Thus, this evidence proves the shortcoming of traditional cutoff-based decisions as well.

(c) Similarity-based methods such as kNN require adequate decision supports from data points nearby. For rare data problem, this kind of method is not able to obtain enough information to distinguish positives from negatives so it is not suitable in application of IMD diagnosis.

In general, most of evaluation results in G_5 are better than G_{16}. It is consistent with the opinion that too few positive samples have impact on predictive accuracy. In our scenario, ML algorithms require approximately 20 positive samples of a disorder to achieve stable performance. With the positive sample grows, models would have higher accuracy for the disease prediction. Among these algorithms, adaboost outperforms all other methods. If a dataset contains only a few positive samples, existing cutoff-based methods or some linear classifiers could be a good choice.

Table 4. Best algorithms for each disease in auxiliary diagnosis.

Disorders	LR	ADA	SVM	LDA	DT
PKU	△	△			
MMA		△			
NICCD	△	*			
SCAD		△		△	
MET		△			
PTPSD	△				
MSUD			△	*	
IVA	△		*		
GA-I	△				
PA				△	
ASS	△	*			
VLCAD	△		*	△	
IBD		*			
GA-II	△			△	
CPT-I			△		△
PRO			△		

Note: The symbol △ denotes a model with at least one highest score on G-mean or F1 metrics, and * indicates a model performs both second best on two metrics. Disorders belonging to G_5 are in bold.

Based on the analysis results, Table 4 lists recommended ML algorithms that achieve relative good predictive performance for each disease. An algorithm is selected if it has one of the highest scores or both second-best in *G-mean* and F_1 metrics. Otherwise, models are omitted in the table. Adaboost is quite appropriate for predicting IMDs if owning enough data of patients. It has preferable

comprehensive ability and performs well in other basic metrics especially in PPV. Besides, linear models cover nearly all diseases so they can be alternatives when lack of positive data.

3.2 Feature Selection

Feature selection methods are designed to automatically filter out irrelevant variables while retaining important features to accomplish certain tasks. Popular feature selection methods are usually as viable options for dataset with tens of thousands of variables, such as gene segments or personalized recommendation, to avoid curse of dimensionality. Although our dataset owns relatively few features, other functionalities of those methods are still helpful for newborn screening data analysis. In this section, we mainly focus on answering *How much performance improvement can a feature selection method brings to predictors* and *Are those selected features reasonable enough to act as a supplement for diagnostic guidance.*

Wrapper-based, filter-based and embedded-based were described as three typical types of feature selection methods [9]. We do not intend to analyze all these methods for the former question, instead, we chose five popular feature selection methods as listed in Table 5. Their selected features were compared to existing diagnostic markers. Other feature construction methods, for instance, principal component analysis (PCA), would not be involved because of interpretable requirements constrained by the latter question.

Table 5. List of five feature selection methods and two baselines compared in this study.

Types	Methods (Abbr.)
Statistics	χ^2 test (Chi2)
	Analysis of variance (ANOVA)
Information theory	Mutual information (MI)
Model oriented	L1-norm (L1-SVM)
	Tree-based (ET)
Baseline	None (origin)
	General standard (existing)

Note: In model-oriented methods, L1-Norm indicates using L1 regularization term to get sparse solution based on SVM classifier and Tree-based uses extremely-randomized tree as the model to fetch important features. We set two baselines: "None" means using all 43 biomarkers without any feature selection and "General Standard" applies metabolic patterns in existing newborn screening.

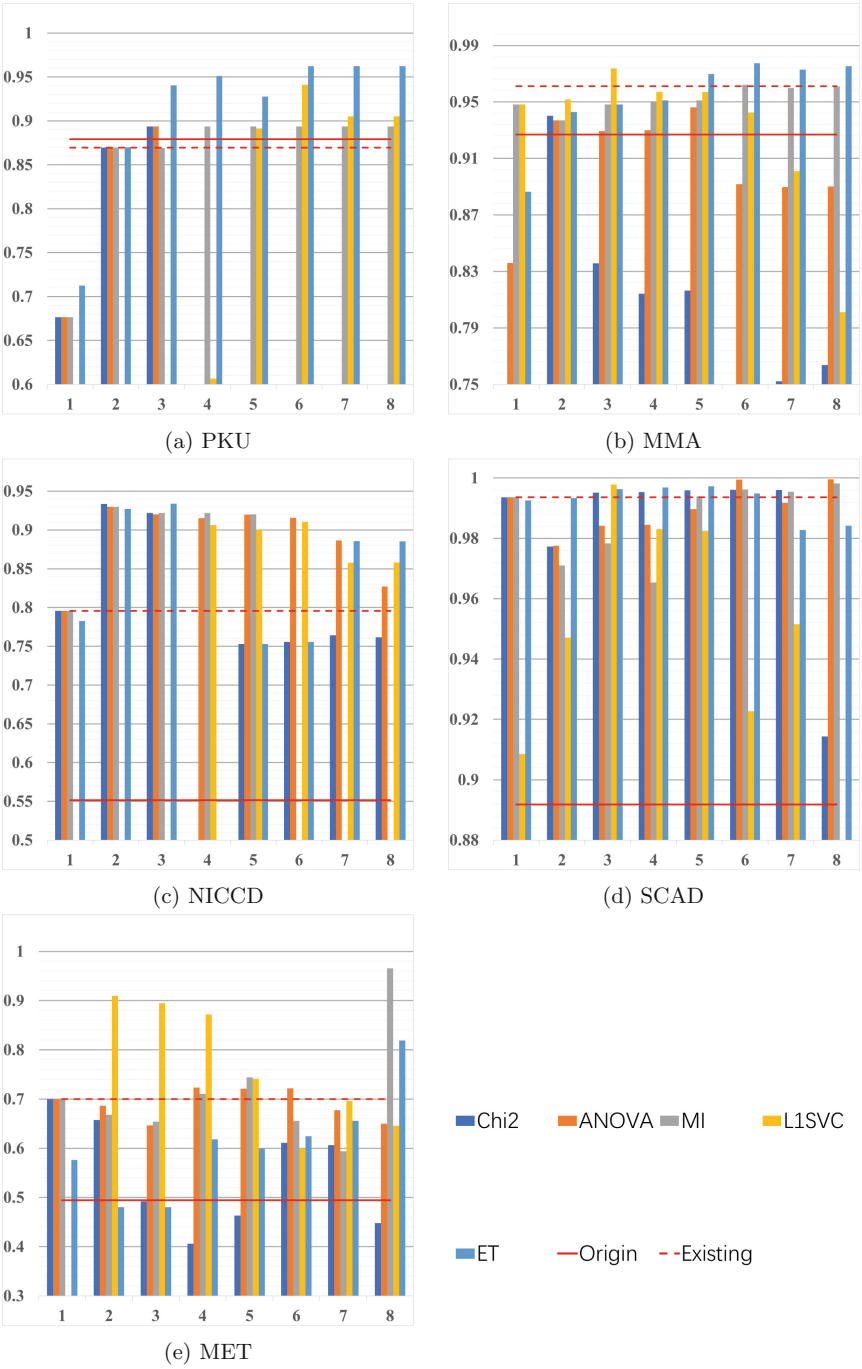

Fig. 2. Predictive performance comparison on G_5 before and after applying feature selection methods and existing diagnostic biomarkers. X-axis represents the number of selected features and Y-axis is average G-mean for all subgraphs.

Figure 2 shows evaluation results of feature selection methods on G_5. Each subgraph consists of two parts: bar graphs are the highest *G-mean* score can be achieved after using different selection algorithms; the dashed line represents the best predictive performance based on existing diagnostic markers and the solid line applies all metabolites in disease prediction. Two lines are treated as baselines to explore effectiveness of these feature selection approaches. Features were selected based on their own criterion such as statistical value or informativeness. In the experiment, we took out the most valuable features considered by different algorithms one at a time and put it into a candidate set $\mathcal{B}$. We iteratively compared changes in prediction performance with the size of $\mathcal{B}$ increase, which are drawn in x-axis as the number of selected features. Up to eight most valuable metabolites are analyzed that account for about 20% of total features. It is remarkable that we enlarge y-axis for better observation of detailed results. Information loss in the figure is inevitable but our primary purpose focuses on higher scores.

From view of differences between two baselines, biomarkers used in newborn screening are truly effective for diagnosis. The results validate the correctness of using existing metabolic patterns. As for feature selection methods, features selected by statistics-based approaches achieve a similar performance to diagnostic markers in general situation. More concretely, ANOVA tends to pick more features, which would not lead to significant improvement or bring any side effects. Chi2 can find out the most relevant features rapidly but is narrowly beaten by ANOVA. Those information theory or model oriented methods greatly outperform statistics-based approaches in some case while keeping similar performance in others. Especially in MET prediction, MI and L1-SVM improve more than 20% performance compared to existing metabolic patterns. Although the first two or three features selected by L1-SVM are usually meaningless, the algorithm can rapidly locate the most valuable features in several more searching steps. Selection criterion based on mutual information performs relatively stable with the size of candidate set changes and the selected metabolic patterns would be useful for diseases prediction. Similarly, the tree-based approach is quite powerful and even the best in the most cases because it involves the idea of both mutual information and bagging in the algorithm. According to the experimental results, model oriented method, especially ET, and mutual information are recommended as auxiliary diagnostic tools for selecting biomarkers.

Table 6. Novel metabolic patterns found by algorithms.

Disorders	Metabolic patterns
PKU	**PHE**, **TYR**, LEU, VAL, PRO, ALA
MMA	C16:1-OH, **C3**, MET, C0, C2, C18:1
NICCD	**CIT**, PHE TYR
SCAD	**C4**, C3, C5:1, C2, SA, C8, C10:1, C6DC
MET	C4, C16:1-OH, C6, C5, PHE, C18:1-OH, C4DC+C5-OH, C18

The most discriminating metabolic patterns found by ML techniques are listed in Table 6. Biomarkers are ranked in the order of their importance. Metabolites are in bold if they are involved in existing diagnostic patterns. For the first four disorders, we can find that these biomarkers are also considered valuable by feature selection algorithms, which verifies correctness of existing patterns. Beyond that, some extra features are deemed to have potential relationships to the cause of disorders. Surprisingly, mutual information approach does not recommend methionine as the top important biomarker to MET. The deeper reason requires further researches by medical experts.

3.3 Rare Data

Many existing ML algorithms assume input data have satisfied the hypothesis of good quality, quantity and representation. Quite the contrary, incompleteness, noise and other data issues in the real application scenarios are completely different from ideal condition. Among these, the class imbalance problem is one of big and common challenges. The term *majority class* denotes the number of samples in a class are overwhelming, otherwise is called *minority class*. There is no definite boundary to distinguish whether a class is considered as majority or not, but in general, the ratio of majority to minority usually reaches hundreds to thousands. Furthermore, rare data problem is a extreme case of imbalance problem, whose ratio is over ten thousand and more. For instance, the majority-minority ratio of our dataset is from 22 thousand to 75 thousand according to data description in Sect. 2. Thus, the absolute quantity of positive samples is rare essentially and it is impractical to solve the insufficiency by increasing overall sample size.

Balancing data by generating simulated samples in minority class or reducing less useful samples in majority class is a natural way to solve this kind of issue. Correspondingly, over-sampling and under-sampling are two types of solutions. We selected two random sampling methods and six advanced works to solve rare data problems, as listed in Table 7. On the basis of metabolic patterns shown in Table 6, re-sampling algorithms are employed to explore possible performance improvements. These algorithms only act on training set without any change in testing set.

Figure 3 shows performance comparisons between before and after applying re-sampling algorithms, which are represented as red lines and bar graphs, respectively. Red lines are used as baselines indicating the highest *G-mean* achieved in Fig. 2. However, it seems these algorithms incur a slight performance degradation in some cases. Two random methods are not stable as expected but still useful in some situation. In general, under-sampling methods perform a little bit better than over-sampling method except One-Sided Selection. But for MET, over-sampling methods are beyond the baseline and achieve higher performance. Core concept of re-sampling techniques utilizes neighbors in the adjacent area to generate simulated samples or drop redundancy. Simulated samples are lack of novelty and informativeness if minor classes provide only a few samples as seeds. Thus, over-sampling methods are not recommended for rare data in

Table 7. List of resampling algorithms for rate data learning.

Types	Methods
Under sampling	Random under sampling
	Tomek links [17]
	One-sided selection [13]
	Edited nearest neighbors [19]
Over sampling	Random over sampling
	SMOTE [6]
	Borderline-SMOTE [10]
Mixed	SMOTE + ENN [1]

general circumstances. However, it could be effective if positive samples have a dense distribution. Generated cases are able to represent relative small sample space with less seed samples. Although under-sampling methods are not helpful for performance improvement, they are also valuable in other applications. For instance, these methods exclude many redundant samples without too much performance degradation, hence selected samples can be used to describe the profile of normal population and new classifiers can be designed based on it.

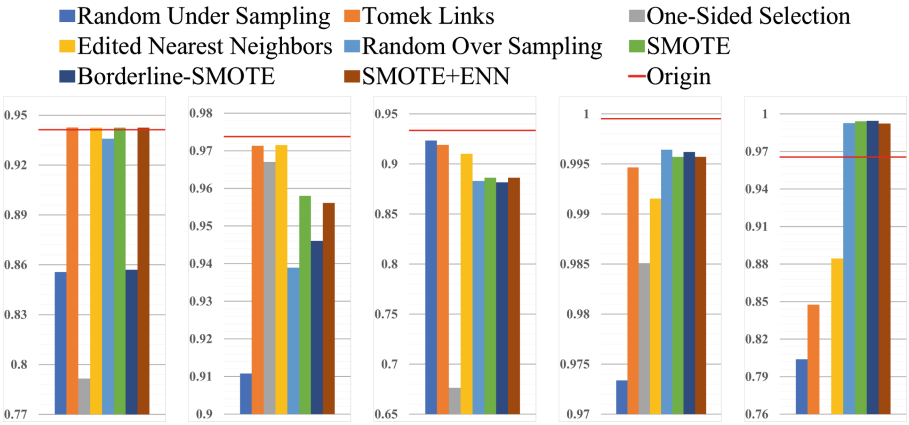

Fig. 3. Predictive performance in G_5 after using re-sampling algorithms. The red line in each subgraph is a baseline representing highest G-mean after applying feature selection algorithms. (Color figure online)

4 Conclusion

Simple cutoff values and doctor's experiences in existing process of newborn screening have limitations on dealing with large-scale MS/MS examination results. To provide more accurate diagnosis of IMDs, we analyze samples of 1.5M neonates and apply several techniques, including ML algorithms, feature selection and re-sampling methods, to improve accuracy in disorder prediction. Experimental results show that adaptive boosting achieves the best comprehensive performance compared to other ML algorithms. Furthermore, feature selection methods are able to find more discriminating metabolites than existing cutoff values on biomarkers. Our analyses also demonstrate that ML algorithms require at least 20 positive samples to achieve stable prediction. For disorders with more than twenty patients, ML techniques can become effective auxiliary diagnostic means for IMDs.

Acknowledgement. This work was supported in part by the National Key Research and Development Program of China (Grant No. 2017YFC1001703, 2018YFC1002700), in part by the National Natural Science Foundation of China (Grant No. 61825205, 61772459) and in part by the National Science and Technology Major Project of China (Grant No. 50-D36B02-9002-16/19).

A MS/MS Biomarkers

Table A1. Overview of 43 biomarkers measured by MS/MS in this study.

Amino acid	Carnitine
Alanine (ALA)	Free (C0)
Arginine (ARG)	Acetyl (C2)
Citrulline (CIT)	Propionyl (C3)
Glutamate (GLU)	Malonyl+Hydroxybutyryl (C3DC+C4OH)
Leucine (LEU)	Butyryl (C4)
Methionine (MET)	Methylmalonyl+Hydroxyisovaleryl (C4DC+C5OH)
Ornithine (ORN)	Isovaleryl (C5)
Phenylalanine (PHE)	Tiglyl (C5:1)
Proline (PRO)	Glutaryl+Hydroxyhexanoyl (C5DC+C6OH)
Tyrosine (TYR)	Hexanoyl (C6)
Valine (VAL)	Methylglutaryl (C6DC)
	Octanoyl (C8)
Ketone	Octenoyl (C8:1)

(*continued*)

Table A1. (*continued*)

Amino acid	Carnitine
Succinylacetone (SA)	Decanoyl (C10)
	Decenoyl (C10:1)
	Decenoyl (C10:2)
	Dodecanoyl (C12)
	Dodecenoyl (C12:1)
	Myristoyl (C14)
	Myristoleyl (C14:1)
	Tetradecadienoyl (C14:2)
	Hydroxytetradecadienoyl (C14OH)
	Hexadecanoyl (C16)
	Hexadecenoyl (C16:1)
	Hydroxypalmitoyl (C16OH)
	Hydroxypalmitoleyl (C16:1OH)
	Octadecanoyl (C18)
	Octadecenoyl (C18:1)
	Linoleoyl (C18:2)
	Hydroxystearoyl (C18OH)
	Hydroxyoleyl (C18:1OH)

References

1. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newslett. **6**(1), 20–29 (2004)
2. Baumgartner, C., Böhm, C., Baumgartner, D.: Modelling of classification rules on metabolic patterns including machine learning and expert knowledge. J. Biomed. Inform. **38**(2), 89–98 (2005)
3. Baumgartner, C., et al.: Supervised machine learning techniques for the classification of metabolic disorders in newborns. Bioinformatics **20**(17), 2985–2996 (2004)
4. Van den Bulcke, T., et al.: Data mining methods for classification of medium-chain Acyl-CoA dehydrogenase deficiency (MCADD) using non-derivatized tandem MS neonatal screening data. J. Biomed. Inform. **44**(2), 319–325 (2011)
5. Chace, D., DiPerna, J., Naylor, E.: Laboratory integration and utilization of tandem mass spectrometry in neonatal screening: a model for clinical mass spectrometry in the next millennium. Acta Paediatr. **88**, 45–47 (1999)
6. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
7. Gulshan, V., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA **316**(22), 2402–2410 (2016)

8. Gurian, E.A., Kinnamon, D.D., Henry, J.J., Waisbren, S.E.: Expanded newborn screening for biochemical disorders: the effect of a false-positive result. Pediatrics **117**(6), 1915–1921 (2006)
9. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003)
10. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) ICIC 2005. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005). https://doi.org/10.1007/11538059_91
11. Hazlett, H.C., et al.: Early brain development in infants at high risk for autism spectrum disorder. Nature **542**(7641), 348 (2017)
12. Iba, W., Langley, P.: Induction of one-level decision trees. In: Machine Learning Proceedings 1992, pp. 233–240. Elsevier (1992)
13. Kubat, M., Matwin, S., et al.: Addressing the curse of imbalanced training sets: one-sided selection. In: International Conference on Machine Learning, Nashville, USA, vol. 97, pp. 179–186 (1997)
14. Lemaître, G., Nogueira, F., Aridas, C.K.: Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. **18**(1), 559–563 (2017)
15. Millington, D., Kodo, N., Norwood, D., Roe, C.: Tandem mass spectrometry: a new method for acylcarnitine profiling with potential for neonatal screening for inborn errors of metabolism. J. Inherit. Metab. Dis. **13**(3), 321–324 (1990)
16. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
17. Tomek, I.: Two modifications of CNN. IEEE Trans. Syst. Man Cybern. **6**, 769–772 (1976)
18. Venditti, L.N., et al.: Newborn screening by tandem mass spectrometry for medium-chain Acyl-CoA dehydrogenase deficiency: a cost-effectiveness analysis. Pediatrics **112**(5), 1005–1015 (2003)
19. Wilson, D.L.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Trans. Syst. Man Cybern. **3**, 408–421 (1972)