Analysis of machine learning algorithms as integrative tools for validation of next generation sequencing data

G. MARCEDDU[1], T. DALLAVILLA[2], G. GUERRI[2], A. ZULIAN[2], C. MARINELLI[2], M. BERTELLI[1]

[1]MAGI Euregio, Bolzano, Italy
[2]MAGI's LAB, Rovereto (TN), Italy

Abstract. – OBJECTIVE: While next generation sequencing (NGS) has become the technology of choice for clinical diagnostics, most genetic laboratories still use Sanger sequencing for orthogonal confirmation of NGS results. Previous studies have shown that when the quality of NGS data is high, most calls are indicated by Sanger sequencing, making confirmation redundant. We aimed at establishing a set of criteria that make it possible to distinguish NGS calls that need orthogonal confirmation from those that do not would significantly decrease the amount of work necessary to reach a diagnosis.

MATERIALS AND METHODS: A data set of 7976 NGS calls confirmed as true or false positive by Sanger sequencing was used to train and test different machine learning (ML) approaches. By varying the size and class balance of the training dataset, we measured the performance of the different algorithms to determine the conditions under which ML is a valid approach for confirming NGS calls in a diagnostic environment.

RESULTS: Our results indicate that machine learning is a valid approach to find variant calls that need more investigation, but in order to reach the high accuracy required in a clinical environment, the training data set must include enough observations and these observations must be well-balanced between true/false positive NGS calls.

CONCLUSIONS: Our results show that it is possible to integrate the diagnostic NGS validation workflow with a machine learning approach to reduce the number of Sanger confirmations of high- quality NGS calls, reducing the time and costs of diagnosis.

Key Words:
NGS, Validation, Diagnostics, Genome analysis, Bioinformatics.

Introduction

Next generation sequencing (NGS) techniques are powerful methods which make it possible to sequence billions of nucleic acid molecules in a single analytical session[1,2]. Today, NGS genetic tests are widely used in the clinical field and have replaced Sanger sequencing (SS)[3], which was previously considered the gold standard for genetic testing. NGS brought a multitude of advantages with respect to SS, for example, multiplexing makes higher throughput possible by sequencing many samples at the same time, and NGS panels can sequence hundreds or thousands of genes simultaneously. NGS also shows lower limits of detection[4] and higher sensitivity in the detection of low-frequency variants[5,6], a fundamental characteristic for the diagnosis of rare genetic diseases and identification of the genes and variants involved. On top of that, NGS techniques dramatically reduce the time and cost of sequencing. Although NGS technology is constantly improving our understanding of genetics, it has several drawbacks. The most significant is that the quality of NGS data can vary considerably, even in the same experiment, depending on a variety of factors, like sequencing technology, target enrichment platform, bioinformatic pipeline, read depth, and mapping accuracy. Even intrinsic properties of the DNA region analyzed can have a great impact on NGS data quality, which can decline for homologous or low complexity/repeated regions. Many genetic laboratories involved in clinical diagnostics therefore choose to use SS for confirmation and validation of NGS results[7], which will be the basis of important clinical decisions. Each variant is therefore confirmed by an orthogonal method to ensure that it is not a false positive. With continuous advances in the accuracy and precision of NGS techniques, it is debated whether laboratories should continue using SS to validate each variant found with NGS. The topic is controversial[7-12]. Different studies have proposed using machine learning (ML) techniques to distinguish

Corresponding Author: Tiziano Dallavilla, MD; e-mail: tiziano.dallavilla@assomagi.org

variants that are confident NGS calls from those requiring SS confirmation. While this proposal seems promising, it can be a challenge to properly train an ML algorithm. The quality of the data used to train an algorithm hugely influences the reliability of its predictions. Moreover, if the validation test is not performed correctly, it may be difficult to realize that the algorithm is not working properly. Machine learning approaches usually require a large training set in which all the classes that the algorithm needs to distinguish are well-represented[13]. This can be a major problem when applying ML to NGS data validation, since collecting numerous NGS calls confirmed by Sanger sequencing is not a trivial task, especially when dealing with rare diseases. Another drawback is that such datasets tend to be highly unbalanced because more than 98% of NGS data is found to be confirmed by SS[9,11,12], so that collecting many false positives may be problematical. Here, we study the effect of different training set parameters on the prediction capacity of different ML algorithms, with the aim of defining good practice for training ML algorithms in the field of NGS validation. First, we discuss the need for Sanger sequencing to validate NGS calls, then we analyze the requirements to implement a ML algorithm for evaluation of variants that need SS confirmation. Our results show that it is possible to develop a ML approach to reliably distinguish true/false positives among high-quality NGS calls, but they also highlight that if this kind of algorithm is not trained and tested properly, it can perform poorly. Moreover, due to the limitations of NGS and the nature of DNA, we conclude that certain types of variant should always be confirmed by SS.

Patients and Methods

Patient Samples and Public Data

A total of 578 patients with rare genetic diseases were enrolled in this study. Depending on their disease, their DNA was sequenced using a targeted NGS approach with custom panels for rare genetic disorders of the eye or for cardiovascular, lymphatic, and metabolic diseases. The first custom panel comprises 234 genes and has a target dimension of 715 kb (CDS ± 15 bp). Mean coverage is 210X and depth of coverage ≥10X, 25X or 40X is 99%, 98.5%, and 97%, respectively. The second custom panel comprises 123 genes and has a target dimension of 370 kb (CDS ± 15 bp). Mean coverage is 340X and depth of coverage ≥10X, 25X or 440X is 98.7%, 98.3%, and 97.6%, respectively. NGS data produced on the two custom panels and obtained with our custom pipeline yielded 1749 potentially clinically relevant variants that were re-sequenced by SS. All the selected variants were rare single nucleotide variants (SNVs), small insertions or small deletions (indels), and rarely synonymous, if already associated with a disease. Sanger sequencing confirmed the presence of 1739 variants (99.43%) and established that 10 variants (0.57%) identified by NGS were actually sequencing artifacts. Another dataset of 7179 NGS calls validated by Sanger sequencing was borrowed from the study of Van den Akker et al[10]. We excluded indels calls, obtaining a refined dataset of 6227 calls, 5754 of which were confirmed to be real calls by SS, whereas 473 variants identified by NGS were sequencing artifacts. These two data sets were pooled to obtain a data set of 7976 calls, 7493 (93.9%) of which were confirmed to be true positives (TPs), while 483 (6.1%) were confirmed to be false positives (FPs). This dataset was then divided into smaller chunks (see below) to train and test the algorithms.

Custom Panel Design

A custom-made oligonucleotide probe library was designed to capture all coding exons and flanking exon/intron boundaries (±15 bp) of genes known from the literature or databases [Human Gene Mutation Database (HGMD Professional), Online Mendelian Inheritance in Man (OMIM), Orphanet, NCBI GeneReviews, NCBI PubMed and specific database] to be associated with the diseases considered. The DNA probe set, complementary to the target regions (GRCh37/hg19), was designed using specific Illumina DesignStudio online tools provided by the company (http://designstudio.illumina.com/Home/SelectAssay/) and optimized with company specialist support to improve the coverage of low-performance target regions.

Panel Design, Library Preparation, and Sequencing

For panel design, library preparation, sequencing, and data analysis by our in-house bioinformatic pipeline we used a workflow already described in our previous papers[14].

Training and Test Datasets

We divided the main dataset of 7976 calls into two subsets. The dataset used for training contained 80% of the data, and that used for testing contained the other 20%. We used the “train test split” function from the Sklearn library to divide the data. The data was first shuffled and then separated into the two datasets in a stratified manner, so that the proportion of TP/FP was maintained in both sets. This operation was repeated three times to obtain three different training and test sets.

Sanger Sequencing

Confirmation of variants identified by NGS covered by 10 reads or more (10X coverage), with a minor allele frequency (MAF) <1% in public databases such as 1000 genome and of likely clinical importance (pathogenic, likely pathogenic, and of unknown significance according to the American College of Medical Genetics and Genomics guidelines[15]) were carried out by bidirectional SS of the target locus with flanking PCR primers designed using the web-based Primer3Plus software[16], avoiding repeated regions and known SNPs in at least the first four bases at 3' of the primer and tested for specificity by primer Blast[17]. Targets were amplified by PCR and underwent agarose gel electrophoresis for size analysis of the resulting amplicons. Unique, properly sized amplicons were purified using standard techniques, while in the case of PCR reactions with unexpected results a second independent set of PCR primers was designed and tested. Sanger sequencing was performed according to the manufacturer's protocols (DTCS starter kit, Absciex) and sequenced on a CEQ8800 Sequencer (Beckman Coulter). Electropherogram analysis was carried out using Chromas (version 2.6.4, Technelysium Pty Ltd).

Results

Selecting Variants for Sanger Sequencing Workflow

Our complete workflow of NGS re-sequencing to detect false positive results is shown in Figure 1. Briefly, variants obtained by NGS data analysis were first evaluated for type, i.e., SNV or indel. This was done because NGS notoriously has problems correctly identifying indels and their positions. It often happens that a single indel can result in multiple calls in close genomic locations, or that the frequency is not estimated correctly[7]. Since there is so much uncertainty about indel calls from NGS we prefer to confirm all indels by SS, thus determining correct genomic position and frequency. The second parameter that we checked was depth: any NGS call not covered at least 10X was confirmed by SS because below 10X, calls show a sharp decline in quality which corresponds to an increased error rate. Finally, we confirmed all variants in highly homologous regions by SS. In these regions call frequency and position may be biased and this may not be reflected by the quality score and other parameters. The other NGS calls were evaluated by the ML algorithm before deciding whether they needed to be confirmed.

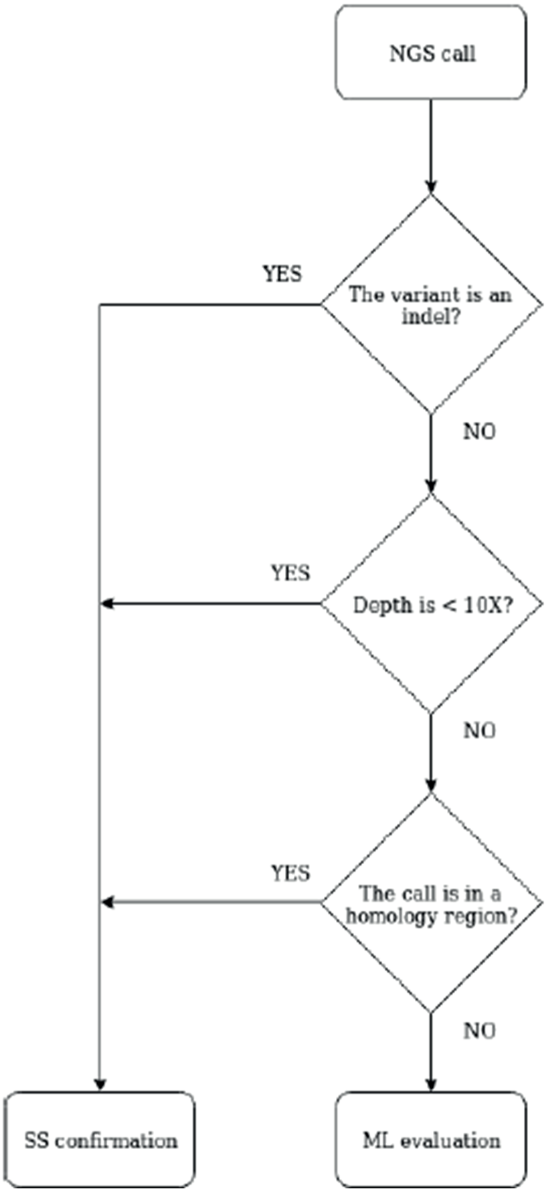

Figure 1. NGS validation workflow. Each NGS call is evaluated before actually deciding whether to perform Sanger sequencing directly or let the ML algorithm decide. If the call is an indel or has a depth <10X or the variant is in a highly homologous region, we confirm it by Sanger sequencing, since the probability of artifacts is high. In all other cases, we evaluate the call by an ML approach to determine whether or not the call needs confirmation.

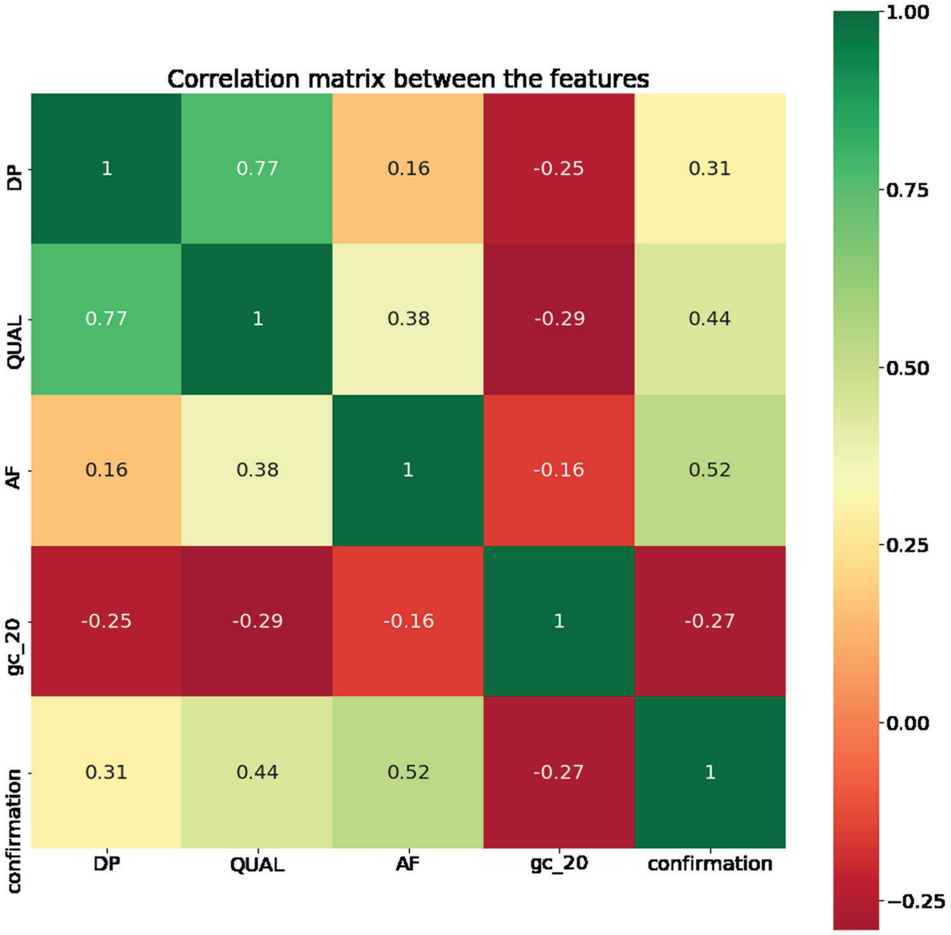

Figure 2. Correlation matrix of the parameters selected for analysis: sequencing depth in call position (DP), score assigned by caller (QUAL), allele frequency (AF), GC content in 20 bases before and after call position (GC 20) and the result of Sanger sequencing (confirmation).

Setting Up a Machine Learning Algorithm

The first step of the analysis was to build a dataset of NGS calls for algorithm training and testing. We selected a number of parameters to describe each call. We tried to pick parameters that were most indicative of call quality, based on our knowledge and observations from previous studies[10,12]. To pick the best parameters for the model, we performed a correlation analysis to find out which correlated best with SS results. To reduce the dimensions of the dataset, we removed parameters with a correlation coefficient <0.25 (Figure 2). The results before filtering are shown in **Supplementary Figure S1**. The features we considered to make predictions were:

- Read depth in call position (Depth)
- Allele frequency (AF)
- GC content in the 20 bases around the call position (GC)
- Standardized quality score from GATK haplotypecaller or Samtools mpileup (QUAL).

In the second step, we selected a number of ML algorithms to test against our dataset, and determined their performance. Since the type of problem we are dealing with can be classified as a supervised learning classification problem, we chose algorithms known to perform well for that type of problem and simple to implement. The algorithms selected were:

- Logistic Regression (LR)
- Nearest Neighbors (NN)
- Linear SVM (LSVM)
- Gradient Boosting Classifier (GBC)
- Decision Tree (DT)
- Random Forest (RF)

Training Test Size

To determine the impact of training set size, each algorithm was trained with a series of datasets built

Table I. Train and Test score along with misclassified FP and TP percentage for training data sets of different sizes.

Algorithm	Train size	Train score	Test score	Misclassified FP (%)	Misclassified TP (%)
Logistic regression	63	0.97	0.95	32.94	2.61
	2552	0.99	0.99	10.79	0.45
	5742	0.99	0.99	9.99	0.35
Nearest neighbors	63	0.94	0.93	100	0.00
	2552	0.97	0.95	51.73	1.97
	5742	0.97	0.97	34.12	0.99
Linear SVM	63	0.97	0.94	97.21	0.02
	2552	0.97	0.95	57.85	1.08
	5742	0.98	0.97	36.31	0.82
Gradient boosting classifier	63	1	0.99	6.09	0.51
	2552	1	0.99	3.9	0.39
	5742	0.99	0.99	1.6	0.23
Decision tree	63	1	0.98	2.79	2.39
	2552	1	0.99	4.98	0.48
	5742	1	0.99	2.14	0.28
Random forest	63	1	0.99	3.3	0.79
	2552	1	0.99	4.98	0.35
	5742	0.99	0.99	0.52	0.21

starting from the training set described in Materials and methods “Training and test datasets”. The original training set was shuffled and divided into subsets of different sizes, ranging from 100% to 10% of the dataset intended for training. Since there was a big difference in the number of FPs and TPs, we decided to stratify when generating the training datasets. This allowed us to maintain the same FP/TP proportion in all the datasets generated, thus minimizing bias from any imbalance. Then, we trained the models with all the training sets and tested their accuracy on three test sets of 1287 NGS calls, each generated from the data intended for testing described in Materials and methods. To assess algorithm performance, we considered the training and test scores along with the percentage of misclassified FPs and TPs. The second parameter is needed to ensure that the algorithm correctly identifies both classes. Since FPs are much fewer than TPs in the test set, it can happen that an algorithm obtains a good overall score solely by correctly identifying TPs. By also checking its capacity to correctly identify FPs, we could determine whether the algorithm was biased by class balance problems during training. Table I and Figure 3 show the performance of the different algorithms for various training set sizes (for complete table see **Supplementary Table S1**). The results show that even if the training and test scores are very high for all the different datasets, the percentage of misclassified FPs and TPs shows a very clear trend of improvement as the training dataset grows in size. All the algorithms benefitted from a larger training dataset. In all cases, we see clearly that the algorithms perform much better at identifying TPs than FPs, revealing that the datasets provided for training were partially unbalanced. However, for the biggest training sets of almost 6000 calls, most algorithms achieved good performance with accuracy exceeding 99.5% in the best case. Only two algorithms showed poor performance: linear SVM and nearest neighbors. While their performance in classifying TPs was still satisfactory, both clearly had difficulty identifying FPs. This test indicates that for medium-small datasets, the algorithms to consider for an ML approach are DT, GBC, and RF. It is possible to train these algorithms with several thousand calls and obtain accurate results. On the other hand, LSVM and NN were clearly penalized by the unbalanced TP/FP ratio and should therefore be avoided for this type of problem with medium-small datasets.

Train Test Balancing

Another major factor that can have a huge impact on the results of training is dataset balance. For proper training, each class (FP and TP in our case)

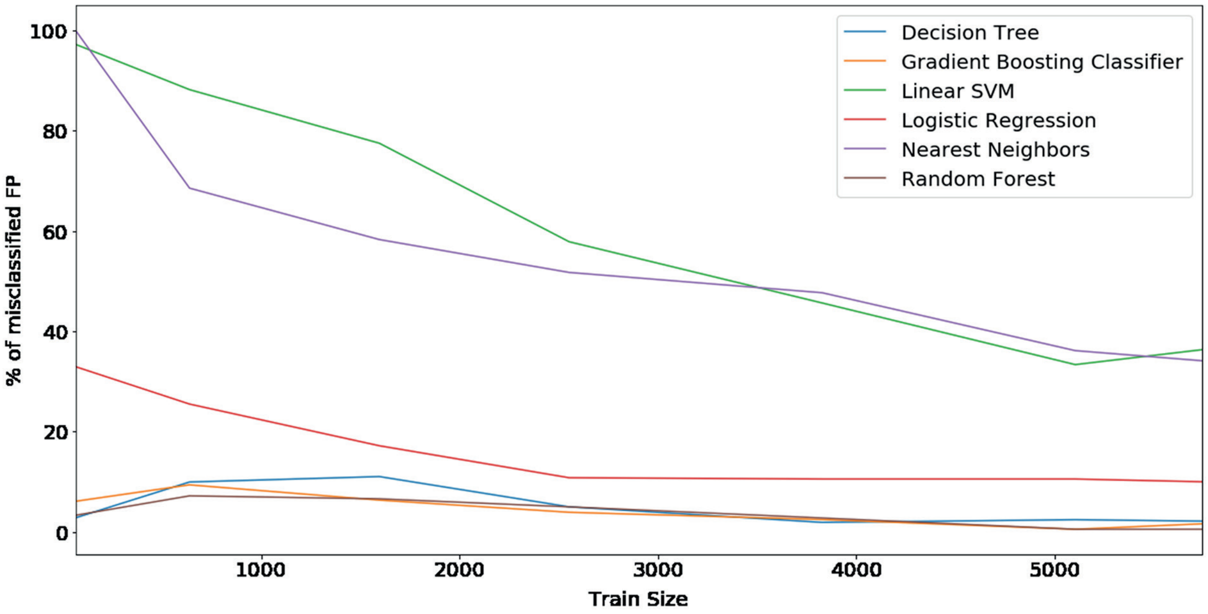

Figure 3. Percentage of incorrectly classified false positives. The graph shows the performance of each algorithm in FP classification for training sets of different sizes.

needs to be properly represented and described. Failing to provide enough data to characterize a class makes it difficult for an algorithm to correctly identify it, which is expressed as poor performance. To understand how the FP/TP balance affected the performance of the different algorithms, we picked the largest dataset for training and used it to create different training datasets, each with a larger percentage of FPs. Dataset size was kept constant while we varied the number of FPs, creating datasets with different balances from FP/TP 1/191 to 1/19. The results are shown in Table II and Figure 4

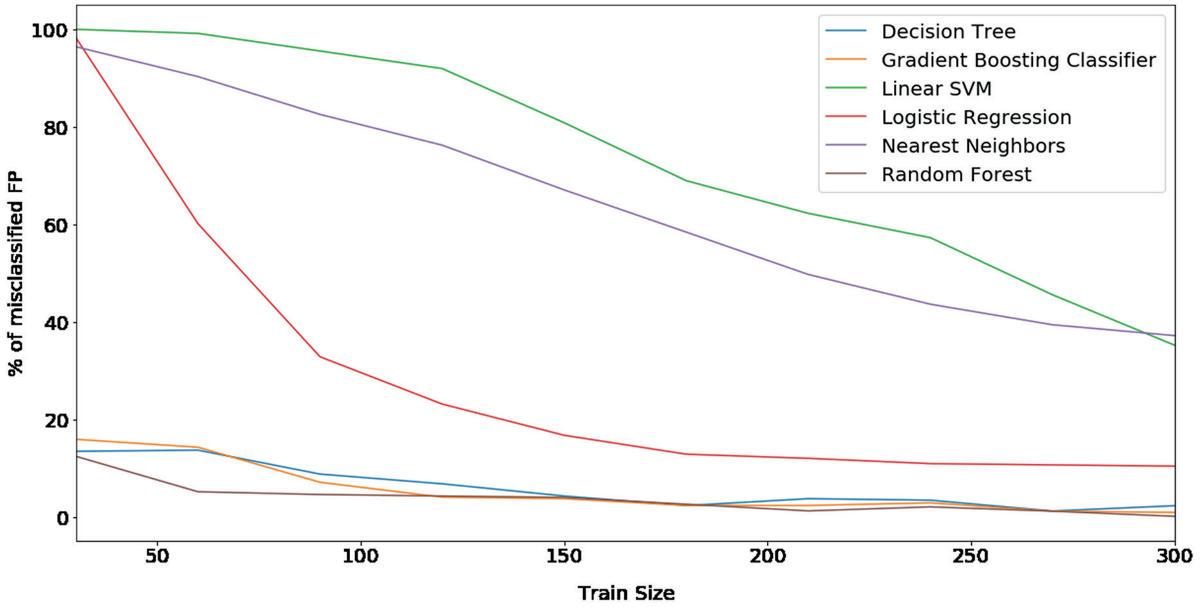

Figure 4. Percentage of incorrectly classified false positives. The graph shows the performance of each algorithm in FP classification for training sets with different FP/TP ratios.

Table II. Train and Test scores and misclassified FP and TP percentages for different FP/TP ratios.

Algorithm	Number of FPs	Train score	Test score	Misclassified FPs (%)	Misclassified TPs (%)
Logistic regression	30	0.99	0.93	98.3	0.02
	180	0.99	0.99	12.98	0.2
	300	0.99	0.99	10.53	0.35
Nearest neighbors	30	0.94	0.93	100	0.00
	180	0.97	0.95	51.73	1.97
	300	0.97	0.97	34.12	0.99
Linear SVM	30	0.99	0.94	96.43	0.00
	180	0.98	0.96	58.47	0.65
	300	0.97	0.96	37.27	1.42
Gradient boosting classifier	30	1	0.99	16.03	0.13
	180	1	0.99	2.48	0.17
	300	0.99	0.99	1.1	0.19
Decision tree	30	1	0.99	13.58	0.13
	180	1	0.99	2.45	0.17
	300	1	0.99	1.34	0.15
Random forest	30	1	0.99	12.55	0.00
	180	1	0.99	2.73	0.23
	300	0.99	0.99	0.26	0.27

(for complete table see Supplementary Table S2). The results indicate that dataset balance is another fundamental property for the proper training of ML algorithms. As the number of FPs increased in the dataset, we saw a huge improvement in the prediction performance of all algorithms. Not all of them were affected by imbalance in the same way: GBC, RF, and DT again seemed less affected, while the prediction performance of LSVM, LM, and LR were clearly more affected by unbalanced datasets. This test again indicates that for medium-small datasets, the best algorithms to consider for an ML approach are DT, GBC, and RF. In any case, it is clear that the best results can only be achieved by ensuring a training dataset in which each class is correctly represented. From our analysis we can conclude that ML algorithms for this type of problem should not be trained with a class imbalance >1/19. With these settings, Random Forest achieved an accuracy >99.5% in the classification of FPs and TPs.

Testing the Algorithm

The method illustrated above was implemented in our in-house pipeline for clinical diagnosis using a decisional tree. The ML algorithm was trained with the dataset used in this study. Subsequently, for each sample analyzed in our laboratory, the ML algorithm was used to evaluate which variants should be confirmed by SS and which not. Table III shows the results for each variant type. The algorithm suggests that most of the variants do not need SS confirmation, independently from the variant type. Overall preliminary results indicate a reduction of Sanger confirmations of about 62%, which is indeed a good result for the reduction of time and cost of the analysis.

Conclusions

In this paper, we discussed the use of Sanger sequencing to confirm NGS results. Sanger sequencing is still required for confirmation of NGS calls made on low-quality data, or in regions that can be particularly problematical, like homologous or low complexity regions. Indels also need to be confirmed by SS, since their position or the exact variant is often difficult to determine precisely with NGS. In these cases, confirmation with an orthogonal method like SS is of primary importance, since the accuracy of NGS in such conditions tends to drop sharply, often leading to artifact calls. Instead when we are dealing with good-quality NGS

Table II. Train and Test scores and misclassified FP and TP percentages for different FP/TP ratios.

Variant type	No. of selected variants	Sanger	No Sanger (%)
Missense_variant	354	109	245 (69.2%)
Intron_variant	28	12	16 (57.1%)
Splice_region_variant&intron_variant	24	10	14 (58.3%)
Stop_gained	17	13	4 (23.5%)
Frameshift_variant	16	16	0 (0.0%)
FRAGMENT	13	13	0 (0.0%)
3_prime_UTR_variant	8	3	5 (62.5%)
Inframe_deletion	7	5	2 (28.5%)
Missense_variant&splice_region_variant	6	0	6 (100%)
Splice_region_variant&synonymous_variant	4	1	3 (75.0%)
5_prime_UTR_variant	3	0	3 (100%)
Splice_donor_variant	3	2	1 (33.3%)
Inframe_insertion	2	2	0 (0.0%)
Splice_acceptor_variant	2	1	1 (50.0%)
Coding_sequence_variant	1	1	0 (0.0%)
Synonymous_variant	1	0	1 (100%)
Total	489	188	301 (61.6%)

Variants selected for Sanger confirmation based on their classification. The table illustrates the reduction of the Sanger sequencing confirmations after using a machine learning approach for the analysis of the variants. The column 'SANGER' indicates those variants that still required confirmation after machine learning analysis while the column 'NO SANGER' indicates the number of variants that were not confirmed by Sanger sequencing.

data, different studies have shown that it is possible to reduce the number of SS confirmations required, since up to 98% of NGS calls are validated by SS. Since finding which NGS calls need more investigation and which do not is a conventional classification problem, the development of a ML algorithm to solve the issue seems intuitive. For an algorithm to work properly, good training is fundamental. We therefore analyzed the minimum requirements in terms of data set size and balance necessary for different algorithms to achieve the accuracy dictated by the stringent criteria of clinical diagnostics. Study of training size highlighted the importance of having a training dataset big enough to allow the model to correctly classify the NGS calls. Our analysis concluded that several thousand calls are needed to properly train the model; with datasets of this size, different algorithms achieved 98% accuracy of prediction for both TPs and FPs, in the best case above 99.5%. Our results also show the importance of having a balanced dataset, namely one in which the classes to identify occur in a sufficient number of NGS calls. We analyzed the effects of imbalance by creating datasets with increasing numbers of FPs. Our results show that when a class is under-represented, the ability of the algorithm to correctly identify NGS calls belonging to that class falls sharply. If the class is also under-represented in the test set, the training and test scores may not highlight the problem, showing high scores even for poorly trained models. In any case, it is also true that a dataset does not need to be perfectly balanced: in our best case, the FP/TP ratio was 1/19, quite distant from 1/1, although different models correctly classified both FPs and TPs with an accuracy exceeding 99%, in the best case above 99.5%.

Regarding the best algorithm to use, we concluded that for medium-small datasets with some balancing bias, the method of choice can be RF, DT, or GBC. These three algorithms seem the least affected by dataset balance and size. The results clearly show that ML, when properly trained, is a powerful approach that can be integrated in the workflow of NGS call confirmation as an alternative to SS orthogonal confirmation. When applied to high-quality NGS data it can considerably reduce the number of confirmations required with an accuracy that permits its use in clinical diagnostics, leading to a faster and less expensive diagnosis.

Conflict of interest

The authors declare no conflicts of interest.

Data Availability

The data used for the study are included in the text and the supplementary information files.

References

1) Makrythanasis P, Antonarakis S. High-throughput sequencing and rare genetic diseases. Mol Syndromol 2012; 3: 197-203.

2) Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem 2009; 55: 641-658.

3) Sikkema-Raddatz B, Johansson LF, de Boer EN, Almomani R, Boven LG, van den Berg MP, van Spaendonck-Zwarts KY, van Tintelen JP, Sijmons RH, Jongbloed JD, Sinke RJ. Targeted next-generation sequencing can replace Sanger sequencing in clinical diagnostics. Hum Mutat 2013; 34: 1035-1042.

4) Schuster SC. Next-generation sequencing transforms today's biology. Nat Methods 2008; 5: 16-18.

5) Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, Boucher G, Ripke S, Ellinghaus D, Burtt N, Fennell T, Kirby A, Latiano A, Goyette P, Green T, Halfvarson J, Haritunians T, Korn JM, Kuruvilla F, Lagacé C, Neale B, Lo KS, Schumm P, Törkvist L; National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC); United Kingdom Inflammatory Bowel Disease Genetics Consortium; International Inflammatory Bowel Disease Genetics Consortium, Dubinsky MC, Brant SR, Silverberg MS, Duerr RH, Altshuler D, Gabriel S, Lettre G, Franke A, D'Amato M, McGovern DP, Cho JH, Rioux JD, Xavier RJ, Daly MJ. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nat Genet 2011; 43: 1066-1073.

6) Jamuar SS, Lam ATN, Kircher M, D'Gama AM, Wang J, Barry BJ, Zhang X, Hill RS, Partlow JN, Rozzo A, Servattalab S, Mehta BK, Topcu M, Amrom D, Andermann E, Dan B, Parrini E, Guerrini R, Scheffer IE, Berkovic SF, Leventer RJ, Shen Y, Wu BL, Barkovich AJ, Sahin M, Chang BS, Bamshad M, Nickerson DA, Shendure J, Poduri A, Yu TW, Walsh CA. Somatic mutations in cerebral cortical malformations. N Engl J Med 2014; 371: 733-743.

7) Baudhuin LM, Lagerstedt SA, Klee EW, Fadra N, Oglesbee D, Ferber MJ. Confirming variants in next-generation sequencing panel testing by sanger sequencing. J Mol Diagn 2015; 17: 456-461.

8) Mu W, Lu HM, Chen J, Li S, Elliott AM. Sanger confirmation is required to achieve optimal sensitivity and specificity in next- generation sequencing panel testing. J Mol Diagn 2016; 18: 923-932.

9) Beck TF, Mullikin JC; NISC Comparative Sequencing Program, Biesecker LG. Systematic evaluation of sanger validation of next-generation sequencing variants. Clin Chem 2016; 62: 647-654.

10) van den Akker J, Mishne G, Zimmer AD, Zhou AY. A machine learning model to determine the accuracy of variant calls in capture-based next generation sequencing. BMC Genomics 2018; 19: 263.

11) Muzzey D, Kash S, Johnson JI, Melroy LM, Kaleta P, Pierce KA, Ready K, Kang HP, Haas KR. Software-assisted manual review of clinical next-generation sequencing data: an alternative to routine Sanger sequencing confirmation with equivalent results in >15,000 germline DNA screens. J Mol Diagn 2019; 21: 296-306.

12) Lincoln SE, Truty R, Lin CF, Zook JM, Paul J, Ramey VH, Salit M, Rehm HL, Nussbaum RL, Lebo MS. A rigorous interlaboratory examination of the need to confirm next-generation sequencing-detected variants with an orthogonal method in clinical genetic testing. J Mol Diagn 2019; 21: 318-329.

13) Banko M, Brill E. Scaling to very very large corpora for natural language disambiguation. In: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2001; 26-33. doi: 10.3115/1073012.1073017. Available online at: https://www.aclweb.org/anthology/papers/P/P01/P01-1005/.

14) Mattassi R, Manara E, Colombo PG, Manara S, Porcella A, Bruno G, Bruson A, Bertelli M. Variant discovery in patients with mendelian vascular anomalies by next-generation sequencing and their use in patient clinical management. J Vasc Surg 2018; 67: 922-932.

15) Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 2015; 17: 405-424.

16) Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res 2007; 35: W71-74.

17) Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-blast: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 2012; 13: 134.