Next Generation Decision Support Tool for Variant Calling

Mick Zomnir

For

Massachusetts General Hospital's Center for Integrated Diagnostics

(MGH CID)

Variant Calling: Background

- What is variant calling?
- Uses NGS data from cancer patients
  - Massively parallel sequencing processes
  - Rapid, reliable, financially reasonable
  - Challenges of harnessing NGS data
- Why it matters: time, money, and lives
- Bionformatics pipelines streamline the process
  - Start with Variant Call Format (.vcf) files
- Current workflow
  - Pathologist reviews .vcf file
  - Interprets variants
  - Issues a clinical report

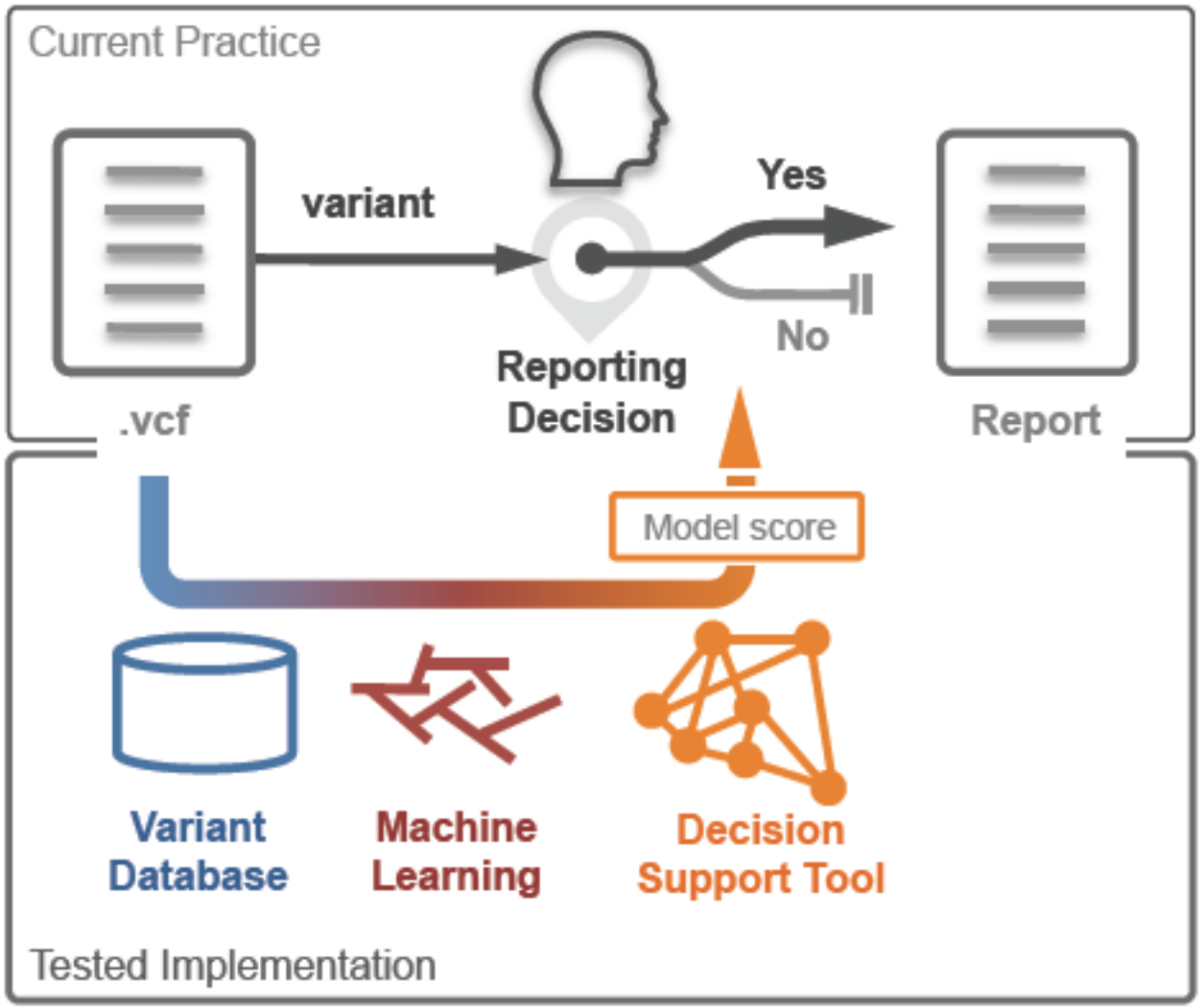
Sample .vcf file

```
##fileformat=VCFv4.1
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order
listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad ma
tes are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as de
fined in the VCF specification">
##GATKCommandLine.HaplotypeCaller=<ID=HaplotypeCaller,Version=3.4-3-gd1ac142,Date="Mon May 18 17:36:4
```

```
[HEADER LINES]
#CHROM  POS ID       REF ALT QUAL    FILTER  INFO          FORMAT          NA12878
1   873762  .        T   G   5231.78 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:173,141:282:99:255,0,255
1   877664  rs3828047   A   G   3931.66 PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  1/1:0,105:94:99:255,255,0
1   899282  rs28548431  C   T   71.77   PASS    [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:1,3:4:26:103,0,26
1   974165  rs9442391   T   C   29.84   LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL  0/1:14,4:14:61:61,0,255
```

Source: Broad Institute GATK website
http://gatkforums.broadinstitute.org/gatk/discussion/1268/what-is-a-vcf-and-how-should-i-interpret-it

Variant Calling: “To Report, or Not To Report”



VarVetter Suite

Variants | CNV | QC Coverage | QC Hotspots | TogetherJS | Counts | FAQs | Undo | Sample QC: PASS

Global Search | 1 selected

VET	ML	CLS	IGV	IMP	MSC	CST	SYM	VSP	VSC	REF	ALT	AF
Search	Min / Max	Search	Search	Search	Search	Search	Search	Search	Search	Min / Max	Min / Max	Min / Max
CHECKED	0.667	SNV	IGV	MODERATE	missense	missense	TP53	p.Asp281Glu	c.843C>A	191	10	0.05
CHECKED	0.533	SNV	IGV	HIGH	stop_gained	stop_gained	APC	p.Arg2204Ter	c.6610C>T	234	20	0.079
CHECKED	0.533	SNV	IGV	MODERATE	missense	missense	TP53	p.Arg282Trp	c.844C>T	130	10	0.071
MAYBE	0.467	SNV	IGV	MODIFIER	upstream_gene	upstream_gene	TERT	null	null	167	11	0.062
YES	0.333	substitution	IGV	MODIFIER	upstream_gene	upstream_gene	TERT	null	null	166	11	0.0625
MAYBE	0.267	SNV	IGV	MODIFIER	upstream_gene	upstream_gene	TERT	null	null	165	11	0.063
YES	0.133	substitution	IGV	MODERATE	missense	missense	TP53	p.AspArg281GluTrp	c.843_844delCCIn	160	10	0.0605
VET	0.000	SNV	IGV	MODERATE	missense	missense	TP53	p.His179Arg	c.536A>G	272	9	0.032
VET	0.000	insertion	IGV	MODERATE	inframe_insertio	inframe_insertio	FGFR1	p.Asp133dup	c.396_398dupTGA	366	2	0.005

The Challenge:

Can we implement a decision support tool, using machine learning, to ~~automate~~ augment the variant calling process?

Data

Data: Overview

- 19,954 variants (all SNPs)
- Assay: NGS-SNAPSHOT
- Sequencing done on Illumina MiSeq
- Cases from November 2013 to June 2016
- Variant calls made by 6 MGH CID molecular pathologists
  - 1 case : 1 pathologist

Table 1. Overview of the Dataset

Component	Number	Description
Variants	**19,954**	**Unique variants from clinical practice**
Disease Sites	**32**	**Unique primary diagnosis sites (e.g. Brain, Lung, etc.)**
Genes	**39**	**Gene names (following HGNC nomenclature)**
Variant Classifications	**4**	**Unique variant classifications (4-tiered system)**
Missense	16,952	Count of Missense Variant
Nonsense	541	Count of Nonsense Variant
Splice Site	228	Count of Splice Site Variant
Start Codon SNP	1,873	Count of Start Codon SNPs
Pathologists	**N=6**	**Number of Pathologists**

Table 2. Number of Cases, Variants, and Calls by Pathologist

Pathologist	Cases	Variants	Variants per case ± SEM	Yes-calls (%)	Yes-calls per case ± SEM	P-values
A	723	4,015	5.55 ± 0.083	959 (23.9)	1.91 ± 0.051	0.45
B	1,118	6,170	5.52 ± 0.064	1502 (24.3)	1.86 ± 0.041	0.73
C	314	1,765	5.62 ± 0.12	419 (23.7)	1.83 ± 0.07	0.38
D	830	4,564	5.50 ± 0.075	1152 (25.2)	1.90 ± 0.041	0.33
E	92	487	5.29 ± 0.20	122 (25.1)	1.77 ± 0.12	0.23
F	453	2,593	5.72 ± 0.099	633 (24.4)	1.97 ± 0.069	0.23
Total	3,530	19,594	5.55 ± 0.036	4787 (24.4)	1.89 ± 0.023	N/A

Abbreviations: SEM, standard error of the mean; N/A, not applicable; P-values derived from Fisher's exact tests comparing each pathologist's average yes-call rate against all others (e.g. A vs. non-A).

Feature	Data type	Data range	Description
Gene	String	See S-Tab. 1	Gene name (following HGNC nomenclature)
Chromosome	Integer	Corresponding to genes listed in S-Tab.1	Chromosome number
Variant Prior	Float	0.0 to 1.0	The frequency of this variant in our dataset
Variant Frequency	Float	0.0 to 1.0	The percentage of alternate reads
Primary Site Diagnosis	String	383 distinct cancer diagnoses	Diagnosis of patient's primary tumor site
Tumor purity in percent	Integer	0* to 100	Pathologist's assessment of percentage of cancerous cells in a specimen
Test Status	String	Abnormal, Equivocal, Fail, Normal, Repeat	Manual quality assessment of the test
Total Reads	Integer	12 to 1,229	Total number of reads per variant site
Alt reads + strand	Integer	0 to 226	Number of alternate reads in positive DNA strand
Alt reads – strand	Integer	0 to 881	Number of alternate reads in negative DNA strand
Ref reads + strand	Integer	0 to 473	Number of reference reads in positive DNA strand
Ref reads – strand	Integer	0 to 850	Number of reference reads in negative DNA strands
Variant Frequency (VF)	Float	0.0 to 1.0	Sum of alternate reads/total reads
Strand bias	Float	0.0 to 1.0	Preference for alternate read fractions between strands is assessed using a Fisher's exact test; a high P-value is indicated of a false positive call.

Feature	Data type	Data range	Description
Variant Classification	String	See Main Table 1	Variant classification in a 4-tiered system
Codon Degeneracy	Integer	0 to 3	Score for the redundancy of the genetic code at the affected codon
Cadd Phred	Float	0 to 55	Score that integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations.
sift	Float	0 to 1	[a]Modified SIFT score [1]
polyphen2_hvar	Float	0 to 1	[a]Modified PolyPhen score [2]
lrt	Float	0.00166 to 0.85682	[a]Modified LRT score [3]
Mutationtaster	Float		[a]Modified MTori score [4]
mutationassessor	Float	0.0 to 1.0	[a]Modified Maori score [5]
Fathmm	Float	0.0 to 1.0	[a]Modified FATHMMori score [6]
provean	Float	0.0 to 1.0	[a]Modified PROVEAN score [7]
vest3	Float	0.0 to 1.0	[a]Modified vest3 score [8]
cadd_raw	Float	0.0 to 1.0	[a]Modified CADD score [9]
metasvm	Float	0.0 to 1.0	[a]Model using multiple orthologous tools to impute a variant prediction [10]
metalr	Float	0.0 to 1.0	[a]More interpretabl derivate model build on metasvm_rankscore [10]
gerp_pp_rs	Float	0.0 to 1.0	[a]Score to account for functional constraints [11]
phylop7way	Float	0.0 to 1.0	[a]Modified phyloP7way score [12]
phastcons7way	Float	0.0 to 1.0	[a]Modifiedl phastCons7way [13]
siphy_29way	Float	0.0 to 1.0	[a]Modified SiPhy_29way_logOdds scores in dbNSFP [14]

Data: Labels and the “Ground Truth”

- Recall: data labels are **binary** (call made by pathologist)
- Model output is **continuous** on 0-1
  - Why?
- Ground truth
  - Did the Steelers win?
  - Did the pathologist report the variant?
  - Was the variant pathogenic?
  - What’s distinct among these questions? What are we *really* modeling?

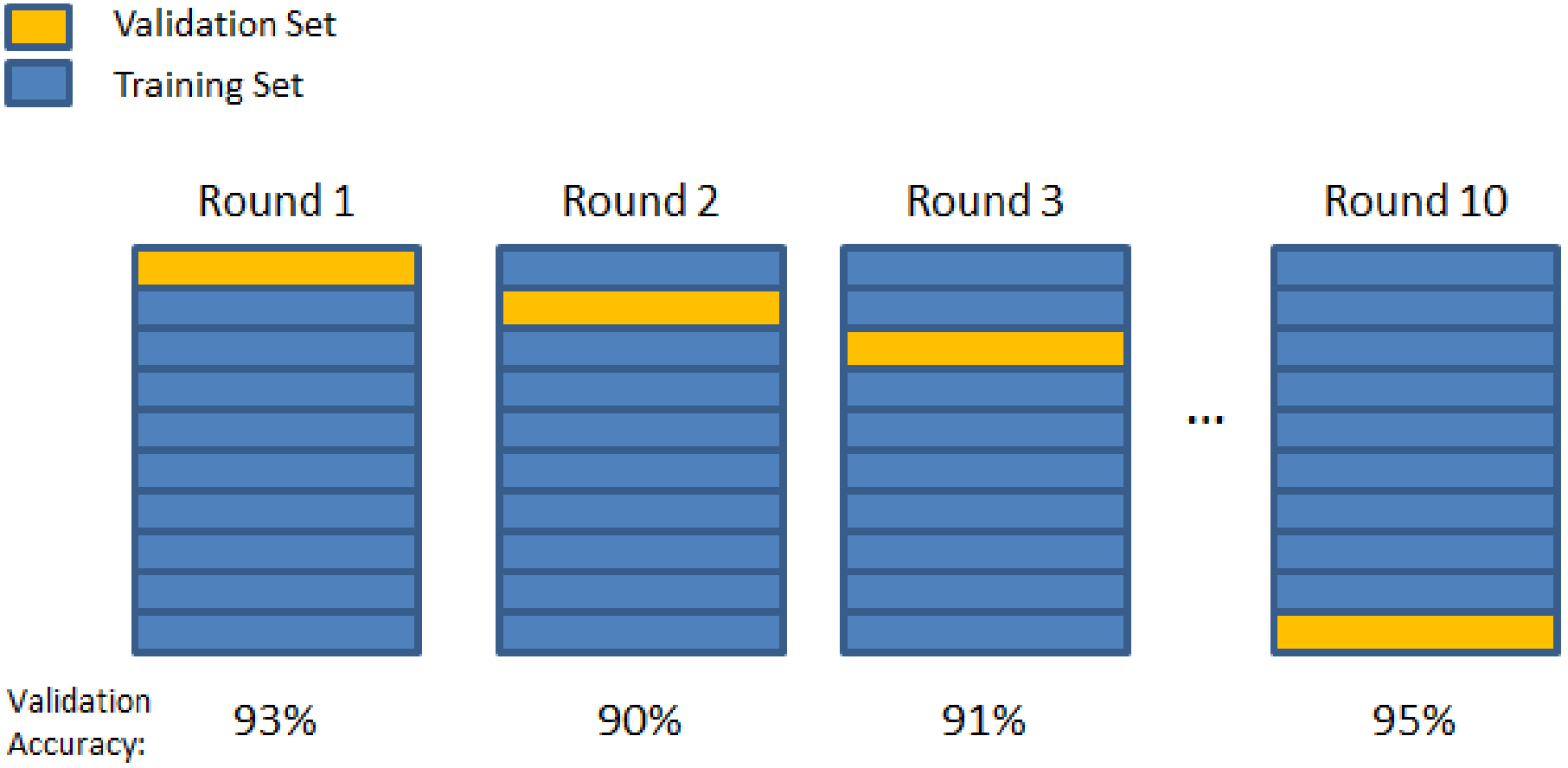
Machine Learning Approach

Machine Learning Approach: Overview

- Used Python programming language
  - Libraries: Pandas, Numpy, scikit learn
- Selected five types of algorithms
  - Naïve Bayes, Decision Trees, Random Forests, SVMs, Logistic Regression
- Performed ANOVA-based feature selection
- Used stratified 10-fold cross validation to compare algorithms
  - Evaluated Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) plot
  - Also considered Precision, Recall/Sensitivity, Specificity, F1 scores
- **Built both Aggregate and Individual Models**

Given a variant, output a prediction (score)

Machine Learning Approach: Cross Validation

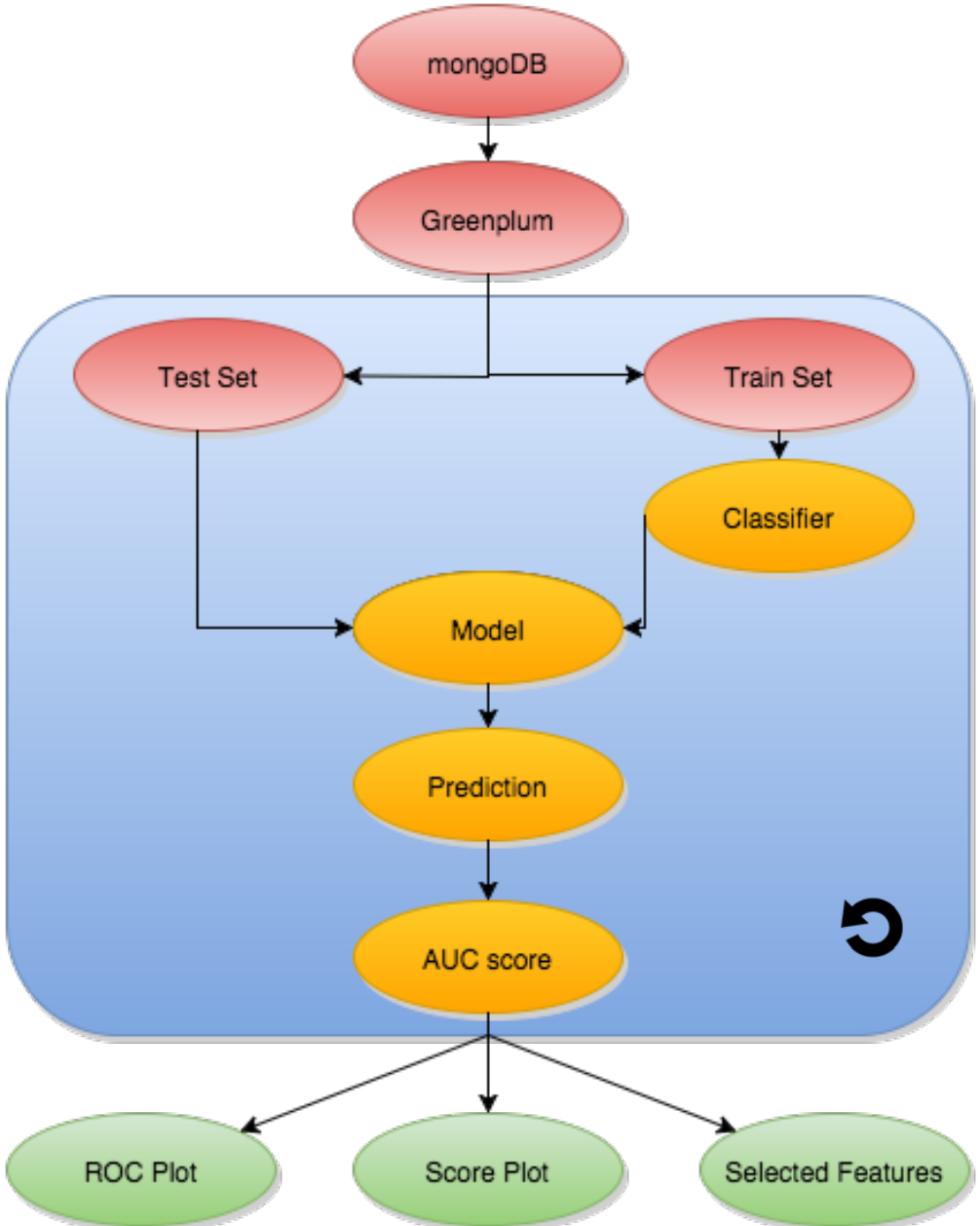


Final Accuracy = Average(Round 1, Round 2, ...)

https://chrisjmccormick.wordpress.com/2013/07/31/k-fold-cross-validation-with-matlab-code/

Machine Learning Approach: Workflow

- Get data from mongoDB and GP
- Train/Test split
- Feed train data into classifier to generate model
- Use model to predict on test set
- Do multiple times (cross validation)
- Output plots/stats



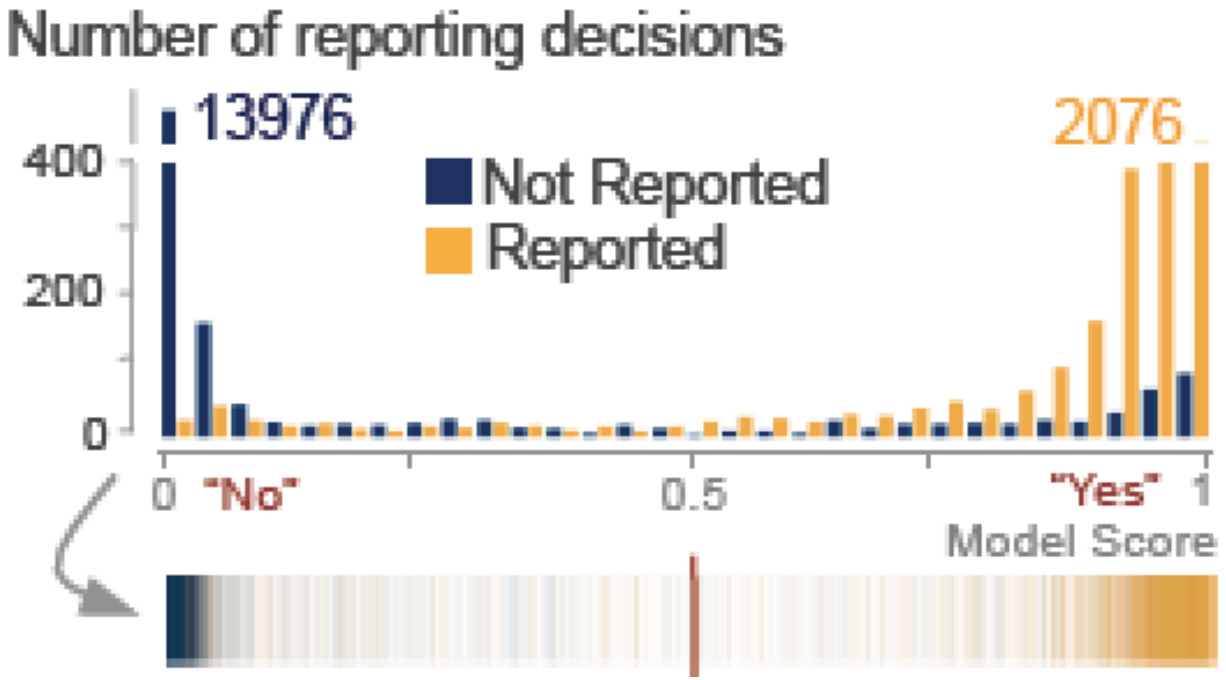
Results

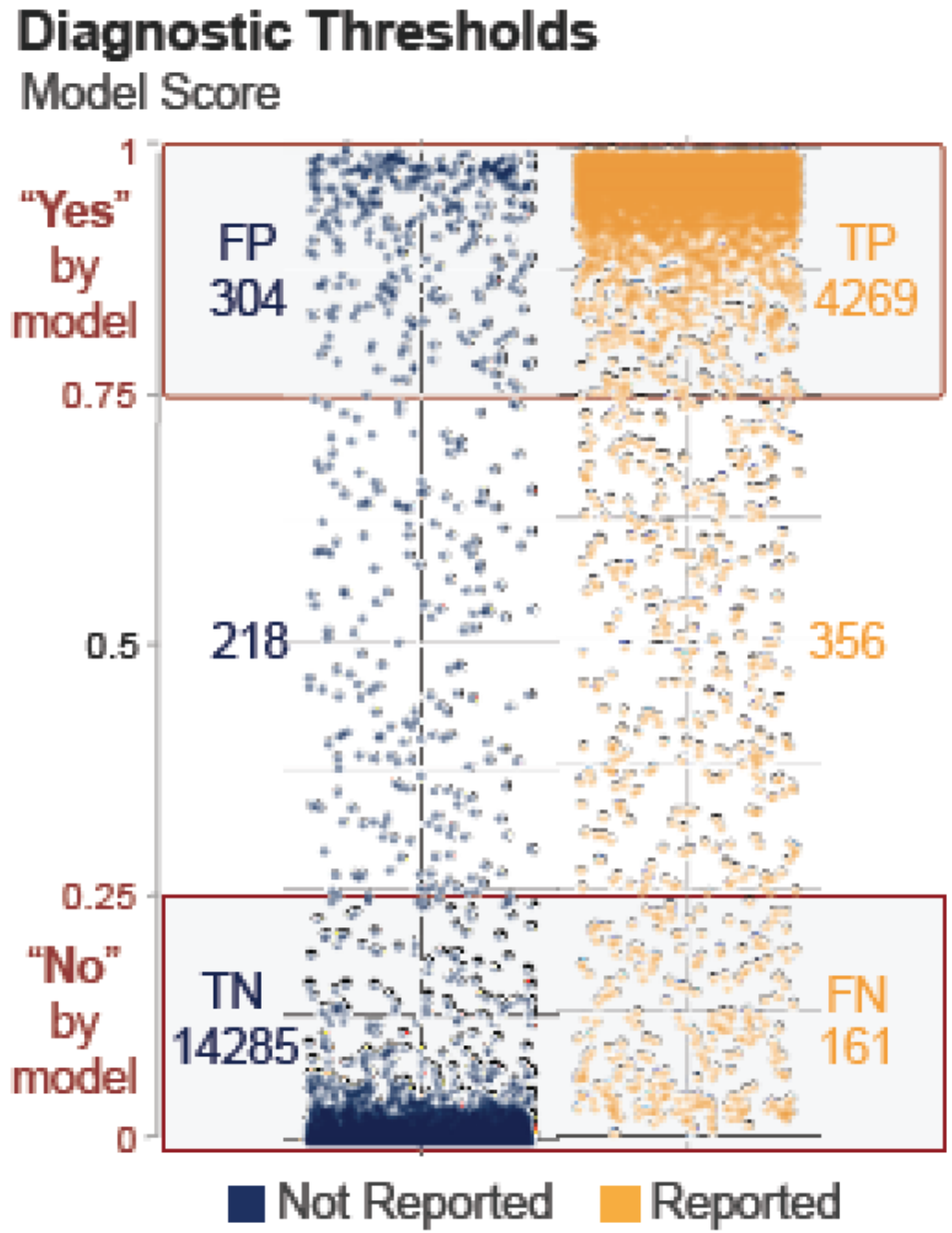
Results: Aggregate Modeling AUCs

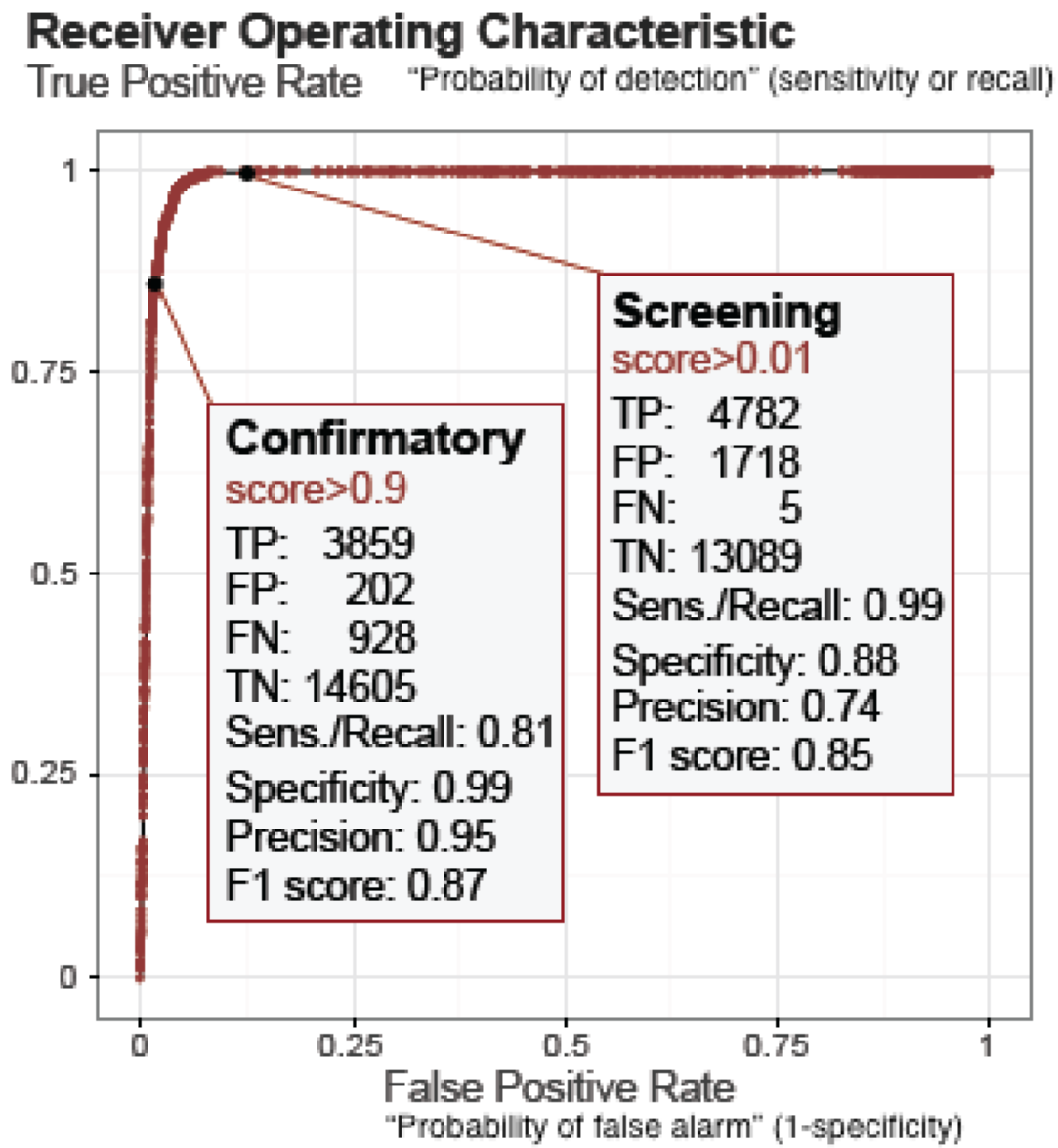
Algorithm	AUC
Random Forests (100 Trees)	0.993
Random Forest (10 Trees)	0.992
Logistic Regression Classifier	0.990
Decision Tree	0.936
Support Vector Machine (rbf kernel function)	0.932
Gaussian Naïve Bayes	0.932

- We chose to implement the Logistic Regression Classifier

Results: Aggregate Modeling (Logistic Regression)




Diagnostic Thresholds
Model Score
1
"Yes" by model
0.75
0.5
0.25
"No" by model
0
FP 304
TP 4269
218
356
TN 14285
FN 161
Not Reported
Reported


Receiver Operating Characteristic
True Positive Rate
"Probability of detection" (sensitivity or recall)
1
0.75
0.5
0.25
0
0
0.25
0.5
0.75
1
False Positive Rate
"Probability of false alarm" (1-specificity)
Confirmatory
score>0.9
TP: 3859
FP: 202
FN: 928
TN: 14605
Sens./Recall: 0.81
Specificity: 0.99
Precision: 0.95
F1 score: 0.87
Screening
score>0.01
TP: 4782
FP: 1718
FN: 5
TN: 13089
Sens./Recall: 0.99
Specificity: 0.88
Precision: 0.74
F1 score: 0.85

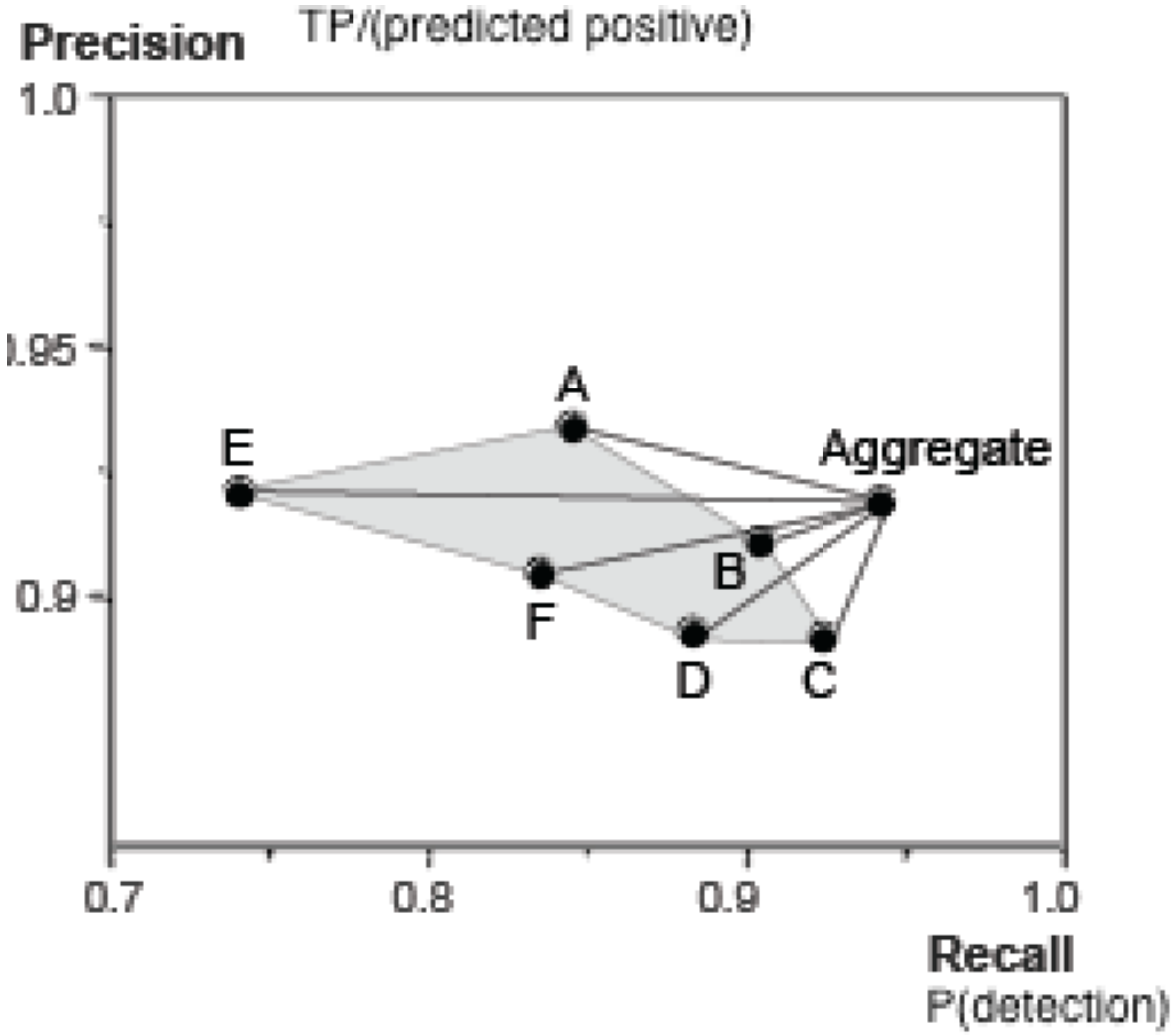
Aggregate Modeling: Top 10 Selected Features

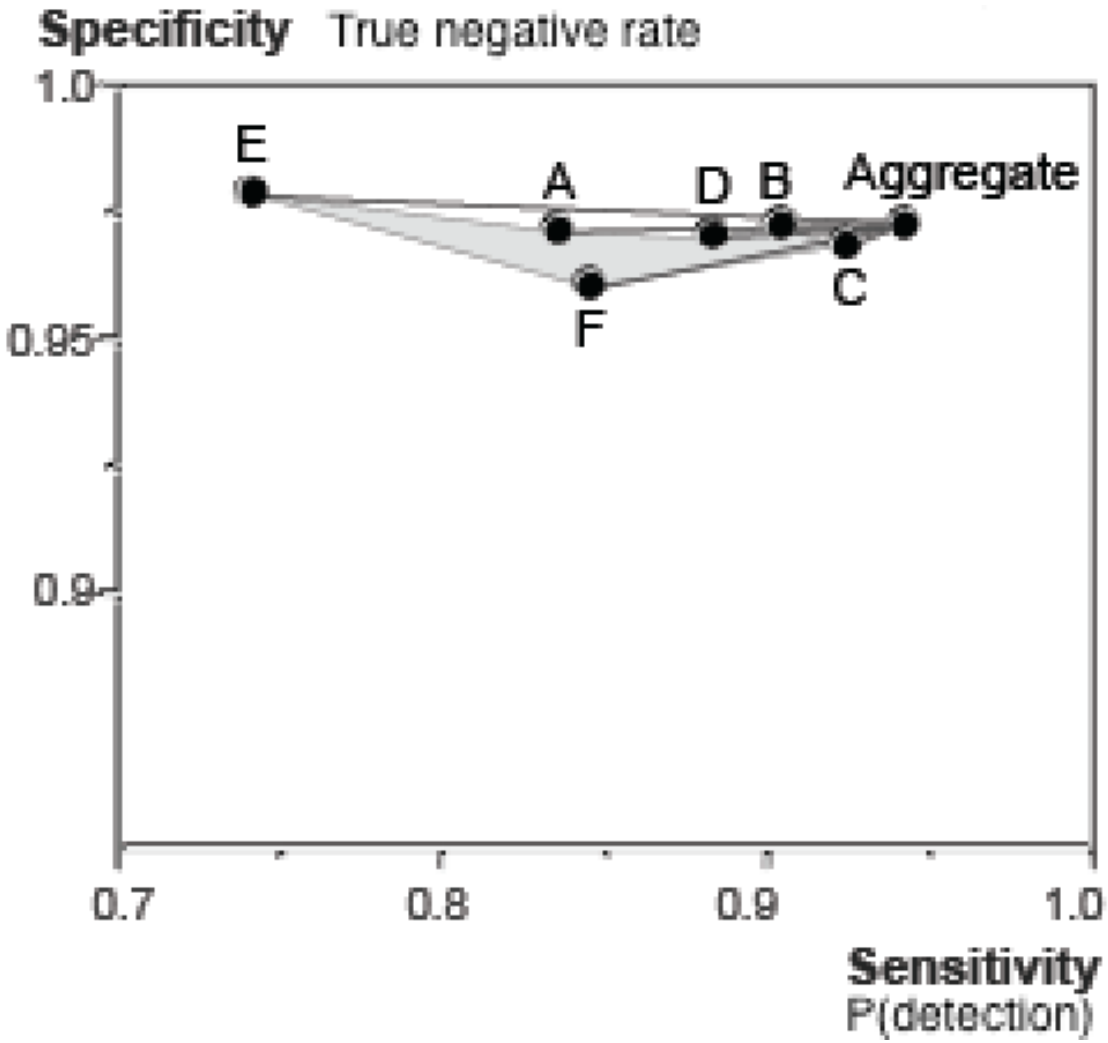
Feature	p-value	Description
metalr_rankscore	0	More interpretable derivate model build on metasvm_rankscore (see below)
genename=KRAS*	0	The gene associated with the variant is/not KRAS
variant_prior	0	The fraction of all variants in our dataset comprised by the given variant
in_hotspot	0	The variant is/not in a hotspot
mutationassessor_rankscore	0	Modified Maori score
vest3_rankscore	0	Modified vest3 score [a]
gerp_pp_rs_rankscore	0	Score to account for functional constraints
metasvm_rankscore	0	Model using multiple orthologous tools to impute a variant prediction
polyphen2_hdiv_rankscore	0	Modified PolyPhen score
n_ref+	0	Count of positive strand alterations found
	0	Modified PROVEAN score [a]

Results: Aggregate and Individual Modeling (Logistic Regression)

Models		Performance Measures				
	TC	AUC	Precision	Recall or Sensitivity	Specificity	F1 Score
Aggregate	**19,594**	**0.990**	**0.919**	**0.942**	**0.973**	**0.930**
Pathologist						
A	4,015	0.986	0.934	0.845	0.961	0.887
B	6,170	0.988	0.911	0.904	0.973	0.907
C	1,765	0.985	0.892	0.824	0.969	0.857
D	4,564	0.988	0.893	0.883	0.971	0.888
E	487	0.982	0.921	0.741	0.979	0.821
F	2,593	0.984	0.905	0.835	0.972	0.868
Transferability						
V1 retrained	568	0.768	0.903	0.567	0.755	0.697
V2 retrained	568	0.980	0.881	0.923	0.959	0.901

Results: Aggregate and Individual Modeling (Logistic Regression)





Toward Clinical Implementation

QC | SNVs | INDELs | COPY NUMBER | SUBMIT

Show 30 entries

Showing 1 to 30 of 52 entries (filtered from 113 total entries)

Search:

First | Previous | 1 | 2 | Next

Report?	VETT		GENE NAME	AA CHANGE	DNA CHANGE	REF	ALT	#TOTAL	#REF_TOT
						Min / Max		Min / Max	
YES			SMAD4	ENSP00000341551:p.Arg416Lys	ENST00000342988.3:c.1247G>A	G	A	110	96
NO			IDH1	ENSP00000390265:p.Met13Lys	ENST00000415913.1:c.38T>A	A	T	52	27
VETT			PIK3R1	ENSP00000428056:p.Gly644Ser	ENST00000521381.1:c.1930G>A	G	A	148	71
VETT			PDGFRA	ENSP00000257290:p.Asp583Gly	ENST00000257290.5:c.1748A>G	A	G	191	178
VETT				00000351276:p.Met391Thr	ENST00000358487.5:c.1172T>C	A	G	182	163
VETT				00000351276:p.Val385Ile	ENST00000358487.5:c.1153G>A	C	T	189	169
VETT				P00000341551:p.Asn3Asp	ENST00000342988.3:c.7A>G	A	G	115	103
VETT	NA	JB		00000341551:p.Arg87Trp	ENST00000342988.3:c.259C>T	C	T	62	50
VETT	NA	JB		P00000308495:p.Val44Ile	ENST00000311936.3:c.130G>A	C	T	183	148
VETT	NA	JB		0000277541:p.Phe2509Ser	ENST00000277541.6:c.7526T>C	A	G	114	91
VETT	NA	JB	NOTCH1	ENSP00000277541:p.Asn1469Asp	ENST00000277541.6:c.4405A>G	T	C	46	34
VETT	NA	JB	EGFR	ENSP00000275493:p.Cys579Arg	ENST00000275493.2:c.1735T>C	T	C	102	89
VETT	NA	JB	NRAS	ENSP00000358548:p.Ser106Pro	ENST00000369535.4:c.316T>C	A	G	119	45
VETT	NA	JB	VHL	ENSP00000256474:p.Val181Ile	ENST00000256474.2:c.541G>A	G	A	127	91
VETT	NA	JB	VHL	ENSP00000256474:p.Met211Thr	ENST00000256474.2:c.632T>C	T	C	111	100
VETT	NA	JB	PIK3CA	ENSP00000263967:p.Leu113Phe	ENST00000263967.3:c.337C>T	C	T	174	158

Model	Report	Score
P1	Yes	93%
P2	No	53%
P3	Yes	95%
P4	Yes	91%
P5	Yes	98%
P6	Yes	99%
Aggregate	Yes	98%

Top 5 Predictors

1. Hotspot
2. Cadd rankscore
3. Variant prior
4. Allelic ratio
5. Mutation rankscore

Summary

- There is a need for a decision support tool to assist pathologists in reporting significant variants in cancer samples.
- This need is addressable with machine learning techniques.
- Deep understanding of both the computational and clinical aspects is crucial for proper implementation.
- Our decision support tool for variant reporting represents one approach to capture the clinical genomics sign-out experience.

MGH CID team

Michael G. Zomnir[1], Maciej Pacula[1], Nishchal Nadhamuni[1], Lev Lipkin[1], Sekhar Duraisamy[1], Enrique Dominguez Meneses[1], Allison MacLeay[1], Saeed H. Al Turki[1*], Zongli Zheng[1], Miguel Rivera[1], Valentina Nardi[1], Dora Dias-Santagata[1], A John Iafrate[1], Long P. Le[1], and Jochen K. Lennerz[1]

Thank you for listening

- Email: michael.zomnir@gmail.com