Supplementary Information

DNN-Boost: Somatic Mutation Identification of Tumor-Only Whole-Exome Sequencing Data Using Deep Neural Network and XGBoost

Firda Aminy Maruf[1], Rian Pratama[1], Giltae Song[1]
[1]School of Computer Science and Engineering, Pusan National University, 63 Busandaehak-Ro, Busan 46241, Republic of Korea

The file includes:

Supplementary 1. Materials & Methods

WES Alignment with Bowtie2

1) Map reads against reference genome:

```
bowtie2 --end-to-end --very-fast --rg-id [ID FOR THE PAIRED-END
READS] -x GRCH38 -q -1 [INPUT FASTQ FILE PAIR 1] -2 [INPUT FASTQ
FILE PAIR 2] | samtools view - -Sb -h -t
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.fai -o [OUTPUT
BAM FILE]
```

2) Sort the output BAM file with SAMTOOLS:

```
samtools sort [INPUT BAM FILE] -o [OUTPUT SORTED BAM FILE] -m
8000000000
```

3) Remove PCR duplicates with SAMTOOLS:

```
samtools rmdup [INPUT SORTED BAM FILE] [OUTPUT SORTED
DEDUPLICATED BAM FILE]
```

Variant Calling

Mutect2

1) Run Mutect2 w/matched normal for the benchmark set

```
gatk --java-options "-Xmx8g" Mutect2 -R
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna -I [INPUT
SORTED DEDUPLICATED TUMOR BAM FILE] -I [INPUT SORTED
DEDUPLICATED NORMAL BAM FILE] -tumor [ID TUMOR BAM FILE] -normal
[ID NORMAL BAM FILE] -pon [PON VCF.gz FILE] --germline-resource
somatic-hg38_af-only-gnomad.hg38.vcf --af-of-alleles-not-in-
resource 0.0000025 -L
Homo_sapiens_assembly38_exome.targets.interval_list -O [OUTPUT
VCF FILE]
```

2) Run FilterMutectCalls to filter somatic variants, germline variants, and artifacts in the Mutect2 VCF callset

```
gatk FilterMutectCalls -R
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna -V [INPUT
MUTECT2 UNFILTERED VCF] -O [OUTPUT MUTECT2 FILTERED VCF]
```

3) Filter out the indels from the Mutect2 filtered VCFs callset

4) Annotate each of the SNP-only VCFs with ANNOVAR to acquire the functional prediction features:

```
perl table_annovar.pl [INPUT MUTECT2 FILTERED VCF] humandb/ -
buildver hg38 -out [OUTPUT ANNOTATED VCF] --remove --protocol
refGene,exac03,avsnp150,dbnsfp33a,gnomad_exome,cosmic92_coding
,clinvar_20210123 --operation gx,f,f,f,f,f,f -nastring . -
vcfinput -polish -xref example/gene_fullxref.txt
```

HaplotypeCaller

1) Run the HaplotypeCaller on each tumor and normal samples BAM files to create single-sample gVCFs, with the option --emitRefConfidence GVCF, and using the .g.vcf extension for the output file.

```
gatk --java-options "-Xmx4g" HaplotypeCaller -R
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna -I [INPUT
SORTED DEDUPLICATED TUMOR BAM FILE] -O [OUTPUT .g.vcf] -A
StrandBiasBySample -ERC GVCF
```

2) Aggregate the multiple GVCF files:

```
gatk --java-options "-Xmx96g -Xms96g" CombineGVCFs -R
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna -V
[INPUT .g.vcf] -V [INPUT .g.vcf] -V [INPUT .g.vcf] -V
[INPUT .g.vcf] -O [OUTPUT FILE COHORT .g.vcf]
```

3) Joint genotyping

```
gatk --java-options "-Xmx4g" GenotypeGVCFs -R
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna -V [INPUT FILE
COHORT .g.vcf] -O [OUTPUT FINAL COHORT VCF]
```

4) Subset to SNPs-only callset with SelectVariants

```
gatk SelectVariants -V [INPUT FINAL COHORT VCF] -select-type SNP
-O [OUTPUT SNP-ONLY VCF]
```

5) Hard-filtering variant

```
gatk VariantFiltration -V [INPUT SNP-ONLY VCF] -filter "QD <
2.0" --filter-name "QD2" -filter "QUAL < 30.0" --filter-name
"QUAL30" -filter "FS > 60.0" --filter-name "FS60" -filter "MQ <
40.0" --filter-name "MQ40" -filter "MQRankSum < -12.5" --filter-
name "MQRankSum-12.5" -O [OUTPUT FILTERED SNP-ONLY VCF]
```

BCFtools

1) Create a list of bams to use:

```
ls *.bam > [OUTPUT BAMLIST .txt]
```

2) Pile the multiple samples, call variants according to the targeted regions, and pipe it to bcftools to create a VCF file:

```
bcftools mpileup -d 250 -R [INPUT TARGETED REGIONS BED FILE] -B
-Ou -f GCA_000001405.15_GRCh38_no_alt_analysis_set.fna -b
[INPUT BAMLIST .txt] | bcftools call -mv -O v -o [OUTPUT VCF]
```

3) Filter query for the variants calling results:

```
bcftools filter -sLowQual -g3 -G10 -e'%QUAL<10 || (RPB<0.1
&& %QUAL<15) || (AC<2 && %QUAL<15)' [INPUT VCF] > [OUTPUT
FILTERED VCF]
```

Supplementary 2. List of Figures

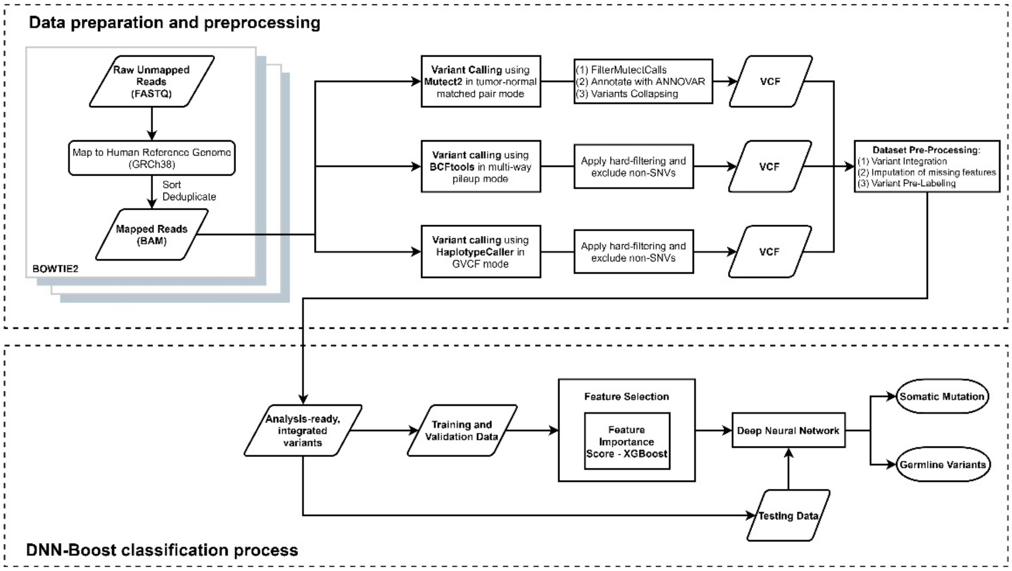
Data preparation and preprocessing
Raw Unmapped Reads (FASTQ)
Map to Human Reference Genome (GRCh38)
Sort
Deduplicate
Mapped Reads (BAM)
BOWTIE2
Variant Calling using Mutect2 in tumor-normal matched pair mode
(1) FilterMutectCalls
(2) Annotate with ANNOVAR
(3) Variants Collapsing
VCF
Variant calling using BCFtools in multi-way pileup mode
Apply hard-filtering and exclude non-SNVs
VCF
Variant calling using HaplotypeCaller in GVCF mode
Apply hard-filtering and exclude non-SNVs
VCF
Dataset Pre-Processing:
(1) Variant Integration
(2) Imputation of missing features
(3) Variant Pre-Labeling
Analysis-ready, integrated variants
Training and Validation Data
Feature Selection
Feature Importance Score - XGBoost
Deep Neural Network
Somatic Mutation
Germline Variants
Testing Data
DNN-Boost classification process

Fig. S1. The overview workflow of the proposed somatic mutation identification pipeline.

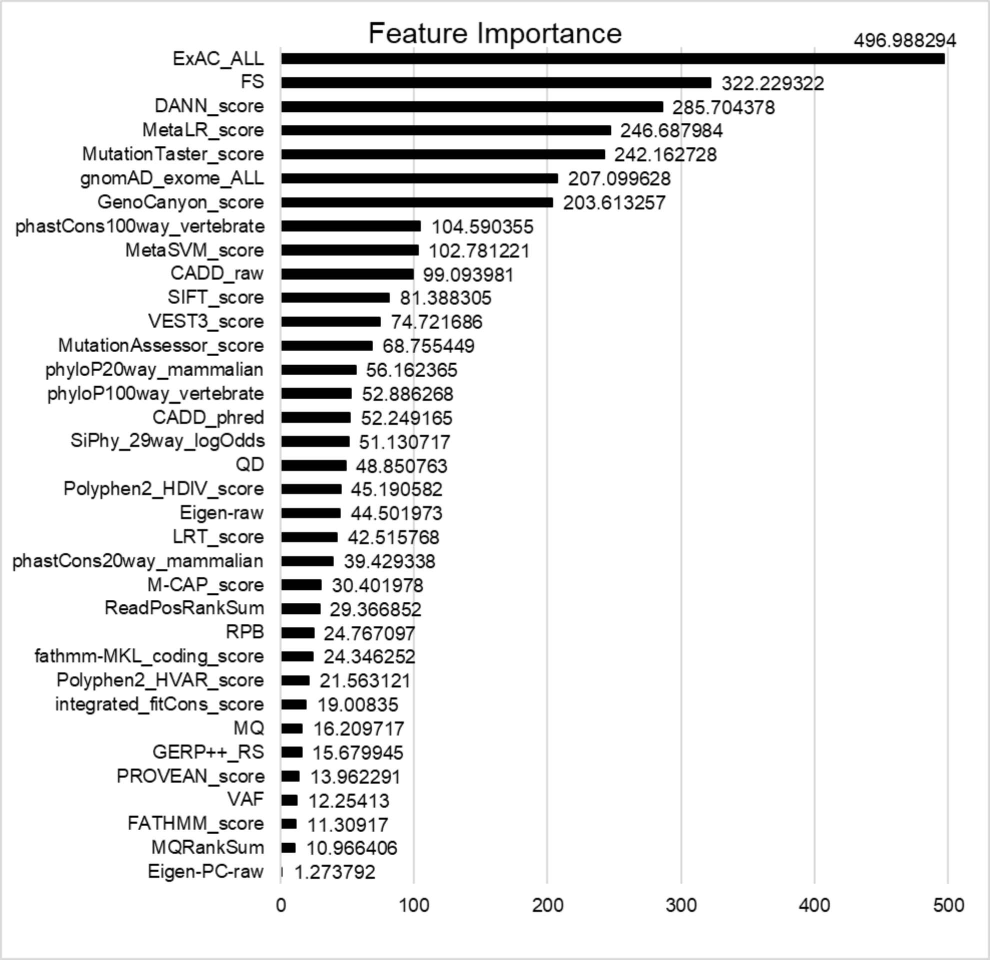

Fig. S2. Feature importance score calculated with gain using XGBoost for the training dataset of paired tumor-normal pancreatic cancer data.

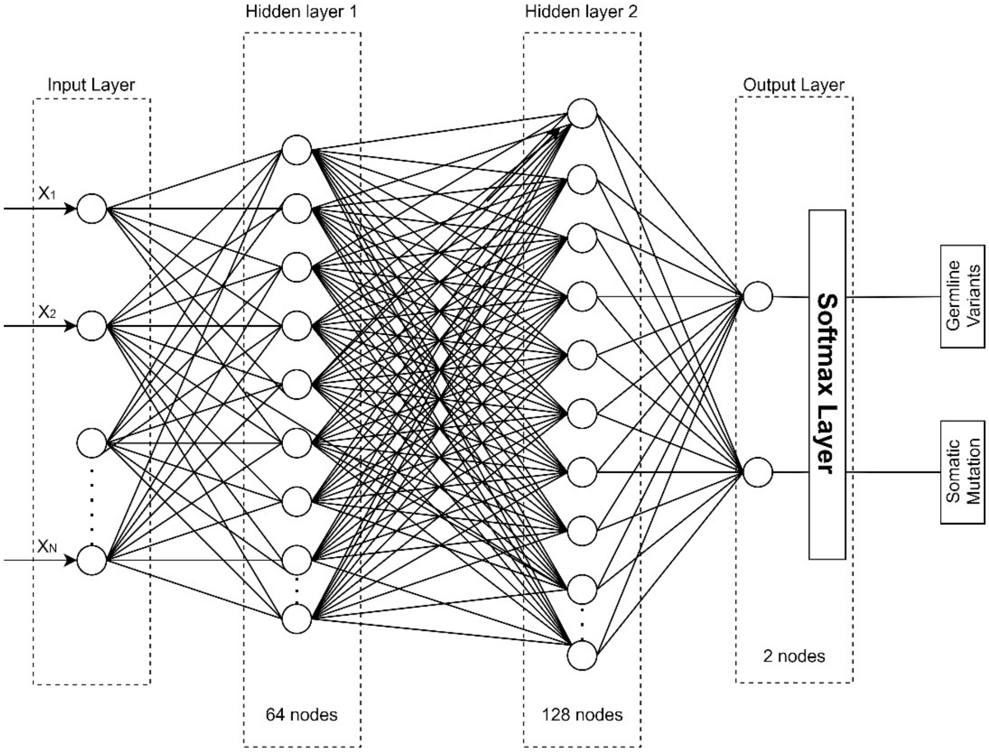

Fig. S3. The proposed architecture of the Deep Neural Network model. The network consisted of four layers, which were an input layer, two hidden layers with 64 and 128 nodes respectively, and an output layer. The number of nodes in the input layer equals to the number of features selected from the XGBoost feature selection method.

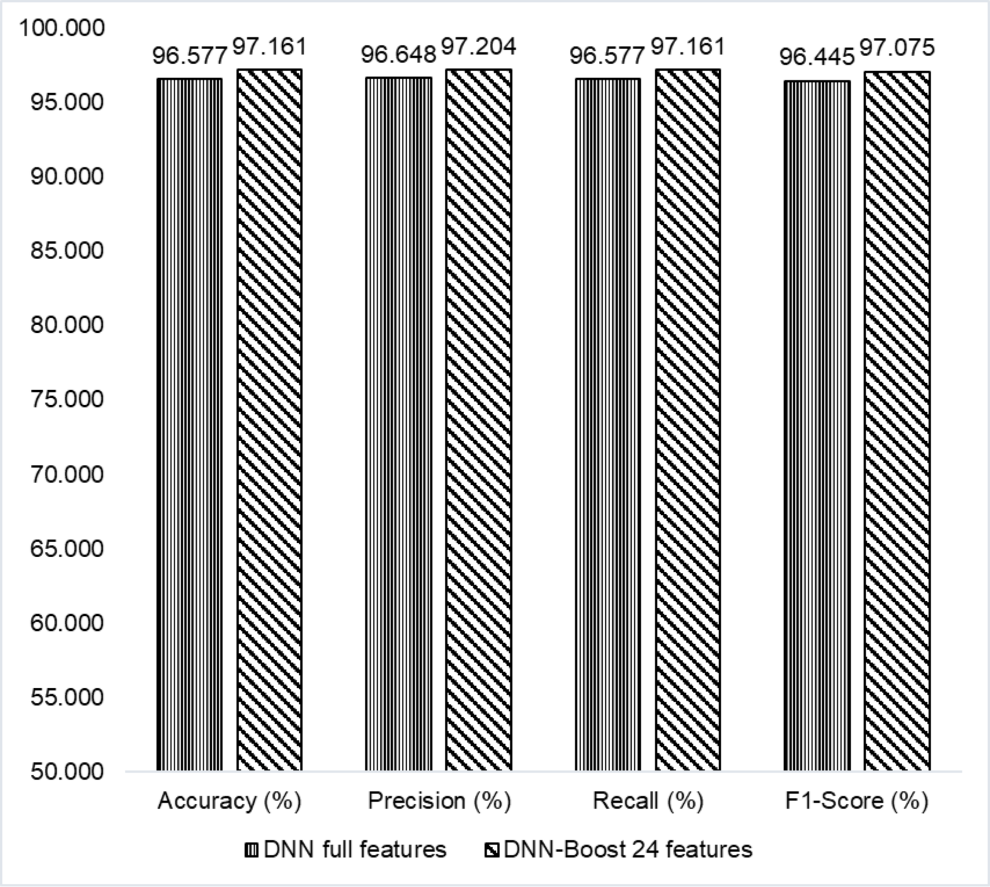

Fig. S4. Performance results of the proposed model with different number of features for classification of paired tumor-normal pancreatic cancer data. The performance measures of DNN-Boost classification of somatic mutation using 24 features achieved the highest accuracy and F1-score of 97.161% and 97.075%, respectively.

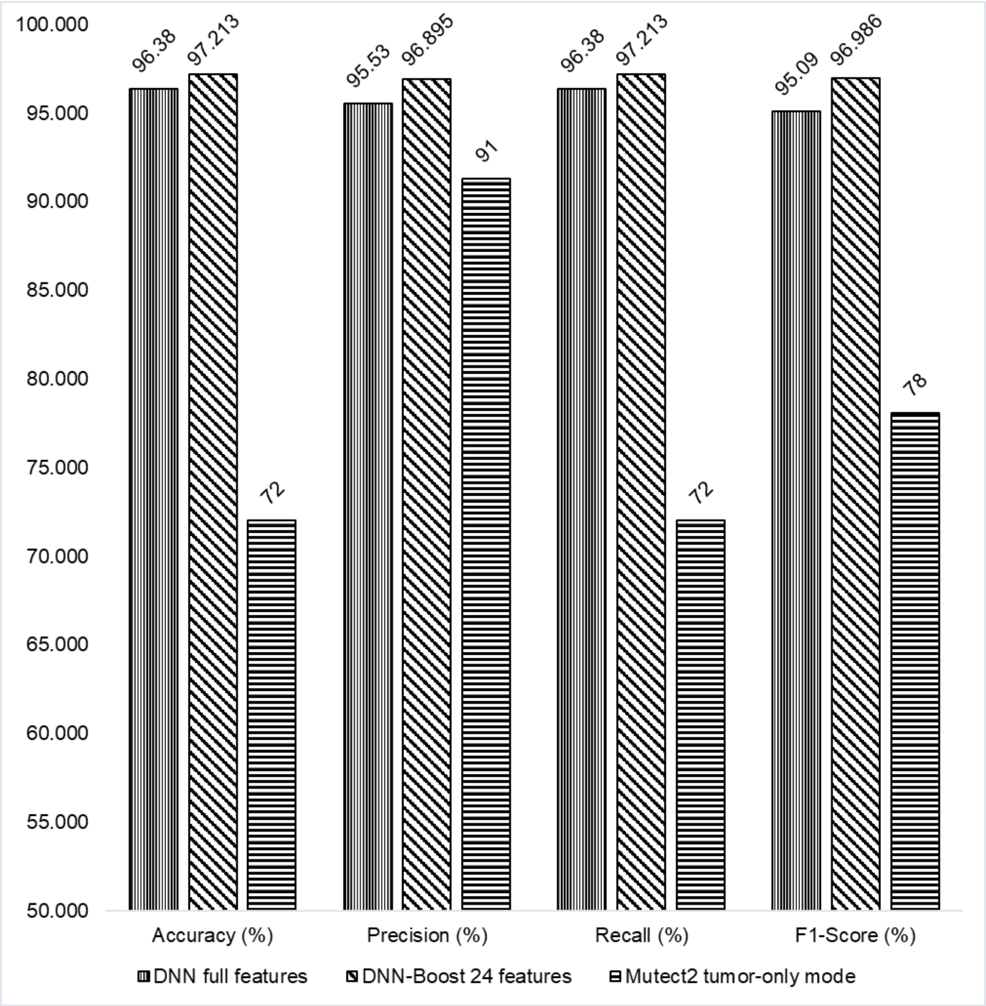

Fig. S5. Comparison of classification performance between the proposed models and the existing method Mutect2 for tumor-only pancreatic samples. The performance measures of DNN-Boost somatic mutation identification of tumor-only pancreatic dataset, using 24 features, achieved the highest accuracy and F1-score of 97.213% and 96.986%, respectively. The Mutect2 tumor-only mode, as the benchmark tool, acquired lower accuracy and F1-score.

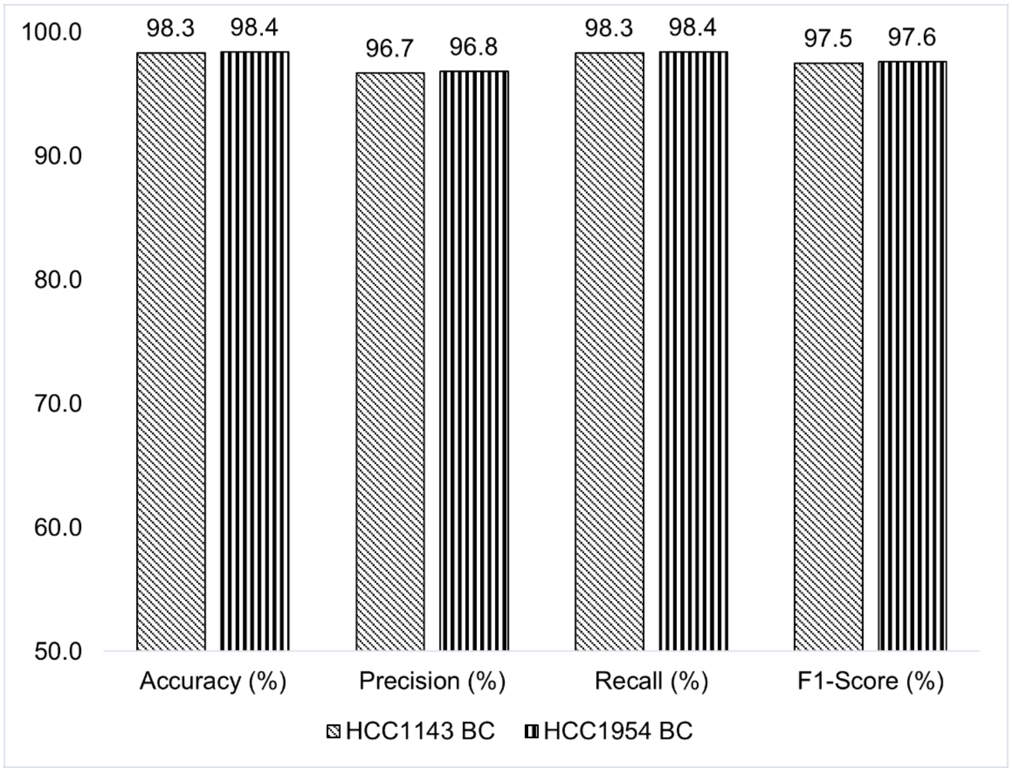

Fig. S6. Performance results of breast cancer (BC) cell line classification with DNN-Boost model using 28 functional prediction features and MQ feature.

Supplementary 3. List of Tables

Table S1. The descriptions of the statistical internal features and variant prediction features used in deep learning classification of somatic mutations.

No	Name	Type	Category
1	Qual By Depth (QD)	Numeric	Statistical internal feature
2	Fisher Strand (FS)	Numeric	Statistical internal feature
3	RMS Mapping Quality (MQ)	Numeric	Statistical internal feature
4	Mapping Quality RankSum Test (MQRankSum)	Numeric	Statistical internal feature
5	ReadPos RankSum Test (ReadPosRankSum)	Numeric	Statistical internal feature
6	Mann-Whitney Test for Differences in Reads (RPB)	Numeric	Statistical internal feature
7	Variant Allele Frequency (VAF)	Numeric	Statistical internal feature
8	The Exome Aggregation Consortium (ExAC) Allele Frequencies	Numeric	Variant allele frequency
9	The Sorting Intolerant from Tolerant (SIFT) Score	Numeric	Functional prediction score
10	The PolyPhen-2 HDIV Score	Numeric	Functional prediction score
11	The Polyphen2 HVAR Score	Numeric	Functional prediction score
12	The Likelihood Ratio Test (LRT) Score	Numeric	Functional prediction score
13	The MutationTaster Score	Numeric	Functional prediction score
14	The MutationAssessor Score	Numeric	Functional prediction score
15	The FATHMM Score	Numeric	Functional prediction score
16	The FATHMM-MKL Coding Score	Numeric	Functional prediction score
17	The PROVEAN Score	Numeric	Functional prediction score
18	The VEST3 Score	Numeric	Functional prediction score
19	The MetaSVM Score	Numeric	Ensemble prediction score
20	The MetaLR Score	Numeric	Ensemble prediction score
21	The M-CAP Score	Numeric	Ensemble prediction score
22	The CADD Raw Score	Numeric	Ensemble prediction score
23	The CADD Phred Score	Numeric	Ensemble prediction score
24	The DANN Score	Numeric	Ensemble prediction score
25	The Eigen Raw Score	Numeric	Functional prediction score
26	The Eigen-PC Raw Score	Numeric	Functional prediction score
27	The GenoCanyon Score	Numeric	Functional prediction score
28	The Integrated FitCons Score	Numeric	Ensemble prediction score
29	The GERP++ RS Score	Numeric	Conservative prediction score
30	The PhyloP100way Vertebrate Score	Numeric	Conservative prediction score
31	The PhyloP20way Mammalian Score	Numeric	Conservative prediction score

Table S1. (*Continued*)

32	The PhastCons100way Vertebrate Score	Numeric	Conservative prediction score
33	The PhastCons20way Mammalian Score	Numeric	Conservative prediction score
34	The SiPhy29way Log Odds Score	Numeric	Conservative prediction score
35	The gnomAD Exome Allele Frequency	Numeric	Variant allele frequency

Table S2. The descriptions of the 24 features selected by XGBoost of the paired tumor-normal pancreatic cancer data, ranked in descending by feature importance score.

No	**Name**	**Score**	**Category**
1	ExAC	496.988294	Variant allele frequency
2	FS	322.229322	Statistical internal feature
3	DANN score	285.704378	Ensemble prediction score
4	MetaLR score	246.687984	Ensemble prediction score
5	Mutation Taster score	242.162728	Functional prediction score
6	gnomAD exome	207.099628	Variant allele frequency
7	GenoCanyon score	203.613257	Functional prediction score
8	phastCons100way vertebrate score	104.590355	Conservative prediction score
9	MetaSVM score	102.781221	Ensemble prediction score
10	CADD raw score	99.093981	Ensemble prediction score
11	SIFT score	81.388305	Functional prediction score
12	VEST3 score	74.721686	Functional prediction score
13	MutationAssessor score	68.755449	Functional prediction score
14	phyloP20way mammalian score	56.162365	Conservative prediction score
15	phyloP100way vertebrate score	52.886268	Conservative prediction score
16	CADD phred score	52.249165	Ensemble prediction score
17	SiPhy29way logOdds score	51.130717	Conservative prediction score
18	QD	48.850763	Statistical internal feature
19	Polyphen-2 HDIV score	45.190582	Functional prediction score
20	Eigen raw score	44.501973	Functional prediction score
21	LRT score	42.515768	Functional prediction score
22	phastCons20way mammalian score	39.429338	Conservative prediction score
23	M-CAP score	30.401978	Ensemble prediction score
24	ReadPosRankSum	29.366852	Statistical internal feature

Table S3. The accuracy and thresholds the test using different subset of features selected by XGBoost.

Threshold	Number of features	Accuracy (%)
0	35	98.37
0.001	30	98.41
0.002	26	98.74
0.003	24	99.59
0.004	21	98.57
0.005	20	98.61
0.006	18	98.61
0.007	16	98.29
0.008	15	98.69
0.009	14	98
0.01	13	98.25
0.012	11	98.29
0.019	9	97.76
0.026	8	97.76
0.056	7	97.8
0.076	6	98
0.09	5	95.02
0.092	4	94.37
0.109	3	94.49
0.128	2	93.55
0.278	1	92.08