**Supplemental Documents**

**Table of contents**

**Supplemental Document 1:**

**ICD‑O‑3 Histology codes used to identify pancreatic ductal adenocarcinoma (PDAC)**

**Supplemental Document 2:**

**Definitions of patient clinical characteristics**

**Supplemental Document 3:**

**Method of pancreas segmentation enhancement**

**Supplemental Document 4:**

**111 Quantitative imaging features and the category they belong to**

**Supplemental Document 5:**

**Hyperparameter setup for support vector machine (SVM) classification model**

**Supplemental Document 6:**

**Exploratory analyses (only for patients without chronic pancreatitis)**

**Supplemental Document 7:**

**Indications for index scans and associated reasons for visit in cases and controls (1 or more) in patients with and without pancreatitis**

**Supplemental Document 8:**

**Detailed Results based on Neighborhood Component Analysis (NCA)**

**Supplemental Document 9:**

**Detailed Results based on Principal Component Analysis (PCA)**

**Supplemental Document 1: ICD‑O‑3 histology codes used to identify pancreatic ductal adenocarcinoma (PDAC)**

8000

8001

8010

8020

8021

8022

8140

8141

8143

8210

8211

8230

8255

8500

8501

8503

8504

8507

8508

8521

8552

8560

8570

8571

8572

8573

8574

8575

8576

Please refer to the Web site for information on the histology codes.

[https://www.naaccr.org/wp-content/uploads/2018/01/](https://www.naaccr.org/wp-content/uploads/2018/01/Updated-Jan-10-2018-ICD-O-3-Guidelines-v2.pdf) [Updated-Jan-10-2018-ICD-O-3-Guidelines-v2.pdf](https://www.naaccr.org/wp-content/uploads/2018/01/Updated-Jan-10-2018-ICD-O-3-Guidelines-v2.pdf).

**Supplemental Document 2: Definitions of patient clinical characteristics**

1. Definition of medical conditions by ICD-9/10 diagnosis codes

|  |  |  |
| --- | --- | --- |
| **Medical conditions** | **ICD-9**  | **ICD-10**  |
| Diabetes  | 250.X 362.0X 357.2 366.41 | E08.X E09.X E10.X E11.X E13.X |
| Biliary tract disease  | 574.X 575.X 576.X 793.3  | K80.X K81.X K82.X K83.X K87.X K91.5 R93.2  |
| Depression  | 311.X 296.0-296.7 296.80 296.82 296.89  | F30.X F31.X F32.X F33.X  |
| Deep vein thrombosis  | 453.0 453.2 453.3 453.4 453.8  | I82.0 I82.210 I82.220 I82.290 I82.3 T82.4 I82.6 I82.A1 I82.B1 I82.C1 I82.890 I82.90  |
| Gallstone disorders  | 574.X 576.1  | K80.X K83.0  |
| Hereditary cancer syndromes | V84.X V16.0 V18.51  | D12.2 D12.3 D12.4 D12.6 Z15.0 Z83.71 Z80.0   |
| Peptic ulcer | 531.x 532.X 533.X 534.X  | K25.X K26.X K27.X K28.X P78.82  |
| Alcohol abuse  | 291.X 303.0 303.9 305.0 357.5 425.5 535.3 571.0 571.2 571.3 760.71 980.0 980.1 E860.1 E860.2 E860.9  | K70.X F10.0 F10.1F F10.2 F10.9 G62.1 G31.2 G72.1 I42.6 K29.2  Q86.0 P04.3 O35.4 K86.0 T51.0 T51.1 T51.9 R78.0   |

1. Definition of weight change in one year
2. Index weight: the weight measure on the index date or closest to and within 6 months prior to the index date.
3. Weight about one year prior to the index date: the weight measure within 9-15 months prior to the index date, and closest in time to (index date – 12 months).
4. Absolute change: the difference between the index weight and the weight about one year prior to the index date.
5. Definition of BMI

BMI= Index weight in kg / (height in m)2

**Supplemental Document 3: Method of pancreas segmentation enhancement**

An algorithm previously developed1 was applied to all the CT scans of cases and controls in datasets DS1 and DS2 to automatically extract the volumetric shape of the pancreas. The process of algorithm enhancement is described below.

Training and validation samples

We took 5 slices of images per CT scan from all eligible PDAC cases. The slices were selected within all the slices involving the pancreas with equal distance from each other. For example, if 18 slices involve the pancreas, the slices 3, 6, 9, 12 and 15 were selected for manual segmentation. Typically, there are 10-80 slices per scan involving pancreas.

For the selected slices, segmentation was manually performed by VA, the imaging specialist. A random sample of 9 scans were selected and manually reviewed by the study radiologist (RP). The annotation was performed for each slice of CT that contained pancreas within the Picture Archiving and Communications System (PACS) dedicated for research.

Process of segmentation enhancement

The cascaded structure was used for extraction of the volumetric shape of the pancreas from axial CT images. The 5 selected slices mentioned above were used as anchor points for enhancing the convolutional neural network. The model parameters derived from the previous study were adjusted to fit the scans of PDAC cases of the current study.2 Adaptive moment estimation was used for adjustment of the weights of the convolutional neural network. The wavelet rendering was applied to extract the volumetric shape from the cross-sectional segmentations.

Performance

To evaluate the performance of the adjusted algorithm, we calculated Dice Similarity Coefficient (DSC) of these 9 randomly selected images of PDAC cases by comparing the automated segmentation and that of manually delineated by the study radiologist.

**Supplemental Document 4: 111 Quantitative imaging features and the category they belong to**

Quantitative features can be broadly categorized into the several subgroups 3: 2D or 3D shape features (e.g. volume, surface area), first-order statistics features (referred to as Global features in the current study), and textual features (second- or higher order statistics). 3 While the first-order statistics features focus on the distribution of individual voxel values such as mean, median and skewness, the second- or higher-order statistics features are calculated considering the inter-relationships between neighboring voxels, and thus capable of detecting heterogeneity (differences) within a specific area. Among these categories, the second-order textual features are known to be most useful for the purpose of differential diagnosis of diseases of the pancreas 4 In the systematic review of radiomics based on pancreas image mining, 100% of studies investigating the relationships between second-order textural features and a pancreas-related outcome found a statistically significant feature 4 To study the utility of quantitative imaging features and pancreatic cancer, Chu et al. applied wavelet and Laplacian transformations on first-order statistics as well as textual features 5

In the current study, we included 16 three-dimensional shape features, 19 first order (Global) features, 76 textual (second- or higher-order statistics) features: 24 Gray Level Co-occurrence Matrix (GLCM) features, 17 Gray Level Run Length Matrix (GLRM) features, 16 Gray Level Size Zone Matrix (GLSZM) features, 14 Gray Level Dependence Matrix (GLDM) features, and 5 Neighboring Gray Tone Difference Matrix (NGTDM) features. All of these 111 features were previously validated and recommended as part of the standardized quantitative radiomics for high-throughput image-based phenotyping. 3 We did not include 10 two-dimensional shape features mentioned in Zwaneburg et al.’s paper because CT images are rendered in three-dimensions.

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Description** | **Feature category** |
| Energy | Energy | Global Feature |
| TotalEnergy | Total Energy | Global Feature |
| Entropy | Entropy | Global Feature |
| Minimum | Minimum | Global Feature |
| TenthPercentile | Tenth Percentile | Global Feature |
| NinetiethPercentile | Ninetieth Percentile | Global Feature |
| Maximum | Maximum | Global Feature |
| Mean | Mean | Global Feature |
| Median | Median | Global Feature |
| InterquartileRange | Interquartile Range | Global Feature |
| Range | Range | Global Feature |
| MAD | Mean Absolute Deviation (MAD) | Global Feature |
| rMAD | Robust Mean Absolute Deviation (rMAD) | Global Feature |
| RMS | Root Mean Squared (RMS) | Global Feature |
| StandardDeviation | Standard Deviation | Global Feature |
| Skewness | Skewness | Global Feature |
| Kurtosis | Kurtosis | Global Feature |
| Variance | Variance | Global Feature |
| Uniformity | Uniformity | Global Feature |
| MeshVolume | Mesh Volume | Shape Features 3D  |
| VoxelVolume | Voxel Volume | Shape Features 3D  |
| SurfaceArea | Surface Area | Shape Features 3D  |
| SurfaceAreatoVolumeRatio | Surface Areato Volume Ratio | Shape Features 3D  |
| Sphericity3D | Sphericity3D | Shape Features 3D  |
| Compactness1 | Compactness1 | Shape Features 3D  |
| Compactness2 | Compactness2 | Shape Features 3D  |
| SphericalDisproportion3D | SphericalDisproportion3D | Shape Features 3D  |
| Maximum3Ddiameter | Maximum3Ddiameter | Shape Features 3D  |
| Maximum2DdiameterSlice | Maximum2DdiameterSlice | Shape Features 3D  |
| Maximum2DdiameterColumn | Maximum2DdiameterColumn | Shape Features 3D  |
| Maximum2DdiameterRow | Maximum2DdiameterRow | Shape Features 3D  |
| MajorAxisLength3D | MajorAxisLength3D | Shape Features 3D  |
| MinorAxisLength3D | MinorAxisLength3D | Shape Features 3D  |
| LeastAxisLength3D | LeastAxisLength3D | Shape Features 3D  |
| Elongation3D | Elongation3D | Shape Features 3D  |
| Flatness | Flatness | Shape Features 3D  |
| AutoCorrelation | Auto Correlation | Gray Level Co-occurrence Matrix (GLCM) |
| JointAverage | Joint Average | Gray Level Co-occurrence Matrix (GLCM)  |
| ClusterProminence | Cluster Prominence | Gray Level Co-occurrence Matrix (GLCM)  |
| ClusterShade | Cluster Shade | Gray Level Co-occurrence Matrix (GLCM)  |
| ClusterTendency | Cluster Tendency | Gray Level Co-occurrence Matrix (GLCM)  |
| Contrast | Contrast | Gray Level Co-occurrence Matrix (GLCM)  |
| Correlation | Correlation | Gray Level Co-occurrence Matrix (GLCM)  |
| DifferenceAverage | Difference Average | Gray Level Co-occurrence Matrix (GLCM)  |
| DifferenceEntropy | Difference Entropy | Gray Level Co-occurrence Matrix (GLCM)  |
| DifferenceVariance | Difference Variance | Gray Level Co-occurrence Matrix (GLCM)  |
| JointEnergy | Joint Energy | Gray Level Co-occurrence Matrix (GLCM)  |
| JointEntropy | Joint Entropy | Gray Level Co-occurrence Matrix (GLCM)  |
| IMC1 | Informational Measure of Correlation (IMC) 1 | Gray Level Co-occurrence Matrix (GLCM)  |
| IMC2 | Informational Measure of Correlation (IMC) 2 | Gray Level Co-occurrence Matrix (GLCM)  |
| IDM | Inverse Difference Moment (IDM) | Gray Level Co-occurrence Matrix (GLCM)  |
| MCC | Maximal Correlation Coefficient (MCC) | Gray Level Co-occurrence Matrix (GLCM)  |
| IDMN | Inverse Difference Moment Normalized (IDMN) | Gray Level Co-occurrence Matrix (GLCM)  |
| ID | Inverse Difference (ID) | Gray Level Co-occurrence Matrix (GLCM)  |
| IDN | Inverse Difference Normalized (IDN) | Gray Level Co-occurrence Matrix (GLCM)  |
| InverseVariance | Inverse Variance | Gray Level Co-occurrence Matrix (GLCM)  |
| MaximumProbability | Maximum Probability | Gray Level Co-occurrence Matrix (GLCM)  |
| SumAverage | Sum Average | Gray Level Co-occurrence Matrix (GLCM)  |
| SumEntropy | Sum Entropy | Gray Level Co-occurrence Matrix (GLCM)  |
| SumofSquares | Sum of Squares | Gray Level Co-occurrence Matrix (GLCM)  |
| SAE | Small Area Emphasis | Gray Level Size Zone Matrix (GLSZM) |
| LAE | Large Area Emphasis  | Gray Level Size Zone Matrix (GLSZM) |
| GLN | Gray Level Non-Uniformity (GLN) | Gray Level Size Zone Matrix (GLSZM) |
| GLNN | Gray Level Non-Uniformity Normalized (GLNN) | Gray Level Size Zone Matrix (GLSZM) |
| SZN | Size-Zone Non-Uniformity (SZN) | Gray Level Size Zone Matrix (GLSZM) |
| SZNN | Size-Zone Non-Uniformity Normalized (SZNN) | Gray Level Size Zone Matrix (GLSZM) |
| ZP | Zone Percentage (ZP) | Gray Level Size Zone Matrix (GLSZM) |
| GLV | Gray Level Variance (GLV) | Gray Level Size Zone Matrix (GLSZM) |
| ZV | Zone Variance | Gray Level Size Zone Matrix (GLSZM) |
| ZE | Zone Entropy (ZE) | Gray Level Size Zone Matrix (GLSZM) |
| LGLZE | Low Gray Level Zone Emphasis (LGLZE) | Gray Level Size Zone Matrix (GLSZM) |
| HGLZE | High Gray Level Zone Emphasis (HGLZE) | Gray Level Size Zone Matrix (GLSZM) |
| SALGLE | Small Area Low Gray Level Emphasis (SALGLE) | Gray Level Size Zone Matrix (GLSZM) |
| SAHGLE | Small Area High Gray Level Emphasis (SAHGLE) | Gray Level Size Zone Matrix (GLSZM) |
| LALGLE | Large Area Low Gray Level Emphasis (LALGLE) | Gray Level Size Zone Matrix (GLSZM) |
| LAHGLE | Large Area High Gray Level Emphasis (LAHGLE) | Gray Level Size Zone Matrix (GLSZM) |
| SRE | Short Run Emphasis (SRE) | Gray Level Run Length Matrix (GLRLM)  |
| LRE | Long Run Emphasis (LRE) | Gray Level Run Length Matrix (GLRLM)  |
| GLNGLRLM | Gray Level Non-Uniformity (GLN) | Gray Level Run Length Matrix (GLRLM)  |
| GLNNGLRLM | Gray Level Non-Uniformity Normalized (GLNN) | Gray Level Run Length Matrix (GLRLM)  |
| RLN | Run Length Non-Uniformity (RLN) | Gray Level Run Length Matrix (GLRLM)  |
| RLNN | Run Length Non-Uniformity Normalized (RLNN) | Gray Level Run Length Matrix (GLRLM)  |
| RP | Run Percentage (RP) | Gray Level Run Length Matrix (GLRLM)  |
| GLVGLRLM | Gray Level Variance (GLV) | Gray Level Run Length Matrix (GLRLM)  |
| RV | Run Variance (RV) | Gray Level Run Length Matrix (GLRLM)  |
| RE | Run Entropy (RE) | Gray Level Run Length Matrix (GLRLM)  |
| LGLRE | Low Gray Level Run Emphasis (LGLRE) | Gray Level Run Length Matrix (GLRLM)  |
| HGLRE | High Gray Level Run Emphasis (HGLRE | Gray Level Run Length Matrix (GLRLM)  |
| LRLGE | Short Run Low Gray Level Emphasis (SRLGLE) | Gray Level Run Length Matrix (GLRLM)  |
| SRHGE | Short Run High Gray Level Emphasis (SRHGLE) | Gray Level Run Length Matrix (GLRLM)  |
| LRLGLE | Long Run Low Gray Level Emphasis (LRLGLE) | Gray Level Run Length Matrix (GLRLM)  |
| LRHGLE | Long Run High Gray Level Emphasis (LRHGLE) | Gray Level Run Length Matrix (GLRLM)  |
| Coarseness | Coarseness | Neighboring Gray Tone Difference Matrix (NGTDM)  |
| ContrastNGTDM | Contrast NGTDM | Neighboring Gray Tone Difference Matrix (NGTDM)  |
| Busyness | Busyness | Neighboring Gray Tone Difference Matrix (NGTDM)  |
| Complexity | Complexity | Neighboring Gray Tone Difference Matrix (NGTDM)  |
| Strength | Strength | Neighboring Gray Tone Difference Matrix (NGTDM)  |
| SDE | Small Dependence Emphasis (SDE) | Gray Level Dependence Matrix (GLDM)  |
| LDE | Large Dependence Emphasis (LDE) | Gray Level Dependence Matrix (GLDM)  |
| GLNGLDM | Gray Level Non-Uniformity (GLN) | Gray Level Dependence Matrix (GLDM)  |
| DN | Dependence Non-Uniformity (DN) | Gray Level Dependence Matrix (GLDM)  |
| DNN | Dependence Non-Uniformity Normalized (DNN) | Gray Level Dependence Matrix (GLDM)  |
| GLVGLDM | Gray Level Variance (GLV) | Gray Level Dependence Matrix (GLDM)  |
| DV | Dependence Variance (DV) | Gray Level Dependence Matrix (GLDM)  |
| DE | Dependence Entropy (DE) | Gray Level Dependence Matrix (GLDM)  |
| LGLE | Low Gray Level Emphasis (LGLE) | Gray Level Dependence Matrix (GLDM)  |
| HGLE | High Gray Level Emphasis (HGLE) | Gray Level Dependence Matrix (GLDM)  |
| SDLGLE | Small Dependence Low Gray Level Emphasis (SDLGLE) | Gray Level Dependence Matrix (GLDM)  |
| SDHGLE | Small Dependence High Gray Level Emphasis (SDHGLE) | Gray Level Dependence Matrix (GLDM)  |
| LDLGLE | Large Dependence Low Gray Level Emphasis (LDLGLE) | Gray Level Dependence Matrix (GLDM)  |
| LDHGLE | Large Dependence High Gray Level Emphasis (LDHGLE) | Gray Level Dependence Matrix (GLDM)  |

1. Asadpour V PR, Mayock RP,Sampson SE,Chen W,Wu BU. Pancreatic Cancer Tumor Analysis in CT Images using Patch-Based Multi-Resolution Convolutional Neural Network. *Biomed Signal Process Control.* 2021;In press.

2. Tajbakhsh N, Shin JY, Gurudu SR, et al. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE transactions on medical imaging.* 2016;35(5):1299-1312.

3. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology.* 2020;295(2):328-338.

4. Abunahel BM, Pontre B, Kumar H, Petrov MS. Pancreas image mining: a systematic review of radiomics. *European radiology.* 2021;31(5):3447-3467.

5. Chu LC, Park S, Kawamoto S, et al. Utility of CT Radiomics Features in Differentiation of Pancreatic Ductal Adenocarcinoma From Normal Pancreatic Tissue. *AJR American journal of roentgenology.* 2019;213(2):349-357.

**Supplemental Document 5: Hyperparameter setup for support vector machine (SVM) classification model**

There are two hyperparameters for SVM. The “Cost” or “C” hyperparameter controls how “hard” or “soft” the margin is. The Gamma hyperparameter controls how much influence individual cases have on the position of the decision boundary. Both hypermeters were tuned based on 5-fold cross-validation. The following tables shows the tuned hyperparameters for the SVM classification models for non-CP patients (training datasets DS1, EDS1, and EDS2 with results reported in Tables 2a, 2b, S2 and S4) and for CP patients (training dataset DS3, results described in texts).

**Non-CP patients:**

|  |  |  |
| --- | --- | --- |
| Training dataset used & result tables | Kernel function | Hyperparameter |
| Cost | Gamma |
| Table 2a (NCA)Training dataset DS1N=414 | Gaussian | 6.3 | 0.251 |
| Linear | 16.6 | 1.77 |
| Sigmoid | 14.3 | 0.0304 |
| Table 2b (PCA)Training dataset DS1N=414 | Gaussian | 5.8 | 0.00365 |
| Linear | 5.8 | 0.00365 |
| Sigmoid | 5.8 | 0.00183 |
| Table S2 (NCA)Training dataset EDS1N=276 | Gaussian | 32.1 | 0.055 |
| Linear | 22.6 | 0.251 |
| Sigmoid | 39.4 | 0.00801 |
| Table S4 (NCA)Training dataset EDS2N=276 | Gaussian | 52 | 0.145 |
| Linear | 47.3 | 0.143 |
| Sigmoid | 49.7 | 0.00418 |

**CP patients:**

|  |  |  |
| --- | --- | --- |
| Training dataset used | Kernel function | Hyperparameter |
| Cost | Gamma |
| NCA Training dataset DS3N=105 | Gaussian | 10.9 | 0.431 |
| Linear | 10.9 | 0.431 |
| Sigmoid | 18.8 | 0.0450 |
| PCA Training dataset DS3N=105 | Gaussian | 13.9 | 0.00173 |
| Linear | 10.9 | 0.431 |
| Sigmoid | 6.48 | 0.0155 |

NCA: for features selected by neighborhood component analysis.

PCA: for features formed by principal component analysis.

**Supplemental Document 6: Exploratory analyses (only for patients without chronic pancreatitis)**

**METHODS**

Because neighborhood component analysis (NCA) method is known to be volatile, we conducted the following exploratory analyses to understand the stability of final classifiers. The entire dataset was randomly split into three equal subsets (EDS1, EDS2, EDS3). We first trained and validated algorithms using EDS1 and EDS3, respectively, based on the same approaches described in the Methods section for the main analyses. We then repeated the process based on EDS2 and EDS3, respectively, for training and validation. The prediction algorithms developed based on the two training datasets (EDS1 and EDS2) were compared using the shared validation dataset (EDS3).

**RESULTS**

**Feature selection**

When EDS1 and EDS2 were used for training, 6 and 14 features were selected by NCA, respectively (Tables S1 and S2). Five of the 6 features selected from EDS1 can be found in that of EDS2. However, the one carrying the largest weight (HGLE) did not appear in the list of features selected from EDS2.

**Performance of developed algorithms**

The performance measures based on the validation data (EDS3) between the two sets of features were similar (Tables S3 and S4). It appears that sensitivity, specificity, PPV, NPV and accuracy were just slightly highly for the algorithm developed based on 14 features when the kernel functions were Gaussian and Sigmoid. However, with a linear kernel function, the classifier based on only 6 features was not worse (or even slightly better) compared to the classifier based on 14 features.

1. **Training based on EDS1 and validation based on EDS3.**

**Table S1. Significant features selected by NCA, the corresponding weights and the category where each feature belongs to.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Order** | **Feature Weight** | **Category** |
| HGLE | 1 | 1.37875717 | GLDM |
| Busyness | 2 | 1.08555260 | NGTDM |
| Strength | 3 | 1.05910526 | NGTDM |
| NinetiethPercentile | 4 | 1.05573412 | Global |
| LDHGLE | 5 | 0.50434021 | GLDM |
| Correlation | 6 | 0.41590827 | GLCM |

Minimum average loss=0.0579 and the corresponding λ=0.0229.

GLDM: Gray Level Dependence Matrix

NGTDM: Neighboring Gray Tone Difference Matrix

GLCM: Gray Level Co-occurrence Matrix

**Table S2. Performance of conditional prediction algorithms with various kernel functions based on the 6 features selected by NCA in Table S1.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Kernel Function** | **Sensitivity (%)** | **Specificity (%)** | **PPV (%)** | **NPV (%)** | **Accuracy (%)** | **AUC** |
| Training: EDS1[3 mos-3 yrs)N=276 | Gaussian | 94.6 | 100.0 | 100.0 | 97.4 | 98.2 | 0.9976 |
| Linear | 93.6 | 97.8 | 95.6 | 96.8 | 96.4 | 0.9959 |
| Sigmoid | 91.3 | 98.4 | 96.6 | 95.8 | 96.0 | 0.9920 |
| Validation: EDS3[3 mos-3yrs)N=279 | Gaussian | 85.0 | 95.2 | 89.8 | 92.7 | 91.8 | 0.9754 |
| Linear | 87.1 | 96.8 | 93.1 | 93.8 | 93.6 | 0.9745 |
| Sigmoid | 81.7 | 97.9 | 95.0 | 91.5 | 92.5 | 0.9833 |

1. **Training based on EDS2 and validation based on EDS3.**

**Table S3. Significant features selected by NCA, the corresponding weights and the category where each feature belongs to.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Order** | **Feature Weight** | **Category** |
| LAHGLE | 1 | 1.391109 | GLSZM |
| NinetiethPercentile | 2 | 1.355935 | Global |
| InverseVariance | 3 | 1.220039 | GLCM |
| Busyness | 4 | 1.159392 | NGTDM |
| Strength | 5 | 0.883597 | NGTDM  |
| Complexity | 6 | 0.869883 | NGTDM |
| ClusterTendency | 7 | 0.749527 | GLCM |
| Correlation | 8 | 0.114777 | GLCM |
| ZP | 9 | 0.100152 | GLSZM |
| LRE | 10 | 0.069448 | GLRLM |
| IDM | 11 | 0.066471 | GLCM |
| LDHGLE | 12 | 0.061537 | GLDM |
| TotalEnergy | 13 | 0.060081 | Global |
| TenthPercentile | 14 | 0.036688 | Global |

Minimum average loss=0.0471 and the corresponding λ=0.0076.

GLSZM: Gray Level Size Zone Matrix

GLCM: Gray Level Co-occurrence Matrix

NGTDM: Neighboring Gray Tone Difference Matrix

GLRLM: Gray Level Run Length Matrix

GLDM: Gray Level Dependence Matrix

**Table S4. Performance of conditional prediction algorithms with various kernel functions based on the 14 selected by NCA in Table S3.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **Kernel Function** | **Sensitivity (%)** | **Specificity (%)** | **PPV (%)** | **NPV (%)** | **Accuracy (%)** | **AUC** |
| Training: EDS2[3 mos-3 yrs)N=276 | Gaussian | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 1.0000.0 |
| Linear | 97.8 | 100.0 | 100.0 | 98.9 | 99.3 | 1.000 |
| Sigmoid | 97.8 | 98.9 | 97.8 | 98.9 | 98.6 | 0.9994 |
| Validation: EDS3[3 mos-3yrs)N=279 | Gaussian | 89.3 | 96.2 | 92.2 | 94.7 | 93.9 | 0.9751 |
| Linear | 87.1 | 94.1 | 88.0 | 93.6 | 91.8 | 0.9774 |
| Sigmoid | 89.3 | 98.4 | 96.5 | 94.8 | 95.3 | 0.9847 |

**Supplemental Document 7: Indications for index scans and associated reasons for visit in cases and controls (1 or more) in patients with and without pancreatitis**

|  |  |  |
| --- | --- | --- |
|  | **Non-chronic pancreatitis** | **Chronic pancreatitis** |
| **Indication for****index scan** | **Cases****N=277** | **Control****N=554** | **Cases****N=70** | **Control****N=140** |
| Abdominal pain | 86 | 131 | 27 | 44 |
| Other pain | 40 | 56 | 3 | 8 |
| Breathing problem | 4 | 16 | 0 | 5 |
| Consultation | 8 | 18 | 1 | 7 |
| Fever | 5 | 8 | 1 | 3 |
| Follow up | 27 | 50 | 1 | 15 |
| GI problem | 39 | 61 | 10 | 28 |
| Other cancer | 5 | 14 | 0 | 4 |
| Test result | 24 | 40 | 6 | 8 |
| Urinary problem | 10 | 26 | 1 | 5 |
| Weakness | 4 | 18 | 0 | 6 |
| Other | 55 | 107 | 14 | 17 |
| Unknown | 33 | 110 | 19 | 24 |
| **Diagnosis** | **Cases****N=277** | **Control****N=554** | **Cases****N=70** | **Control****N=140** |
| Abdominal pain | 57 | 79 | 11 | 23 |
| Other pain | 21 | 28 | 2 | 2 |
| GI problem | 67 | 94 | 25 | 48 |
| Other cancer | 26 | 98 | 4 | 12 |
| Test results | 4 | 6 | 0 | 2 |
| Urinary problem | 32 | 61 | 6 | 7 |
| Weight loss | 12 | 6 | 0 | 0 |
| Other | 84 | 177 | 21 | 51 |
| Unknown | 29 | 56 | 12 | 7 |

**Supplemental Document 8: Detailed results based on Neighborhood Component Analysis (NCA)**

**Non-CP patients:**

The first figure below shows the average loss values versus λ values for the non-CP patients. The minimum average loss was 0.0459, and the corresponding λ (the best) was 0.0178. Since maximum weight was 1.4050, five features were selected because their weights were over 0.02810 (2%\*1.4050). The table below shows the significant features selected by NCA, the corresponding weights and the category where each feature belongs to. None of the five features was dropped during the backward feature elimination process.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Order | Feature Weight | Category |
| LDHGLE | 1 | 1.405048088 | GLDM |
| Busyness | 2 | 1.088298766 | NGTDM |
| NinetiethPercentile | 3 | 1.071671022 | Global  |
| Strength | 4 | 0.884799771 | NGTDM |
| HGLE | 5 | 0.336837041 | GLDM |



The figure below shows the feature weight versus features index under the best λ, 0.0178. 5 significant features were selected using the relative threshold of 0.02810.



**CP patients:**

The first figure below shows the average loss values versus λ values for the CP model. The minimum average loss was 0. There was a range of λ with average loss of 0. We selected λ=0.0019 (smallest) as “the best” for the table below and further analyses. Since maximum weight was 1.9634, fourteen features were selected because their weights were over 0.03927 (2%\*1.9634). The table below shows the significant features selected by NCA, the corresponding weights, and the category where each feature belongs to.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | Order | Feature Weight | Category |
| JointAverage | 1 | 1.963430672 | GLCM |
| AutoCorrelation | 2 | 1.707044044 | GLCM |
| LGLRE | 3 | 0.361684907 | GLRLM  |
| SumAverage | 4 | 0.176525317 | GLCM |
| HGLZE | 5 | 0.120574151 | GLSZM |
| Maximum3Ddiameter | 6 | 0.104239105 | Shape Features 3D |
| SphericalDisproportion3D | 7 | 0.096952004 | Shape Features 3D |
| MinorAxisLength3D | 8 | 0.086462798 | Shape Features 3D |
| GLV | 9 | 0.074624987 | GLSZM |
| SAHGLE | 10 | 0.064103658 | GLSZM |
| Contrast | 11 | 0.055467685 | GLCM |
| GLVGLRLM | 12 | 0.052715689 | GLRLM |
| RP | 13 | 0.040656497 | GLRLM |
| IDMN | 14 | 0.040606340 | GLCM |



The figure below shows the feature weight versus features index under the best λ, 0.0019. 14 significant features were selected using the relative threshold of 0.03927.



During the backward feature elimination process, we dropped the feature with the lowest feature weight (an iterative process) if the reduction of AUC was less than 0.001. Only the top 2 features remained at the end of the selection process. We investigated additional choices of lambda (not shown). For all investigated lambda, the top 2 features (JointAverage and AutoCorrelation) were selected.

**Supplemental Document 9: Detailed results based on Principal Component Analysis (PCA)**

The original 111 quantitative imaging features were transformed into principal components.

**Non-CP patients:**

The top 25 with the highest standard deviation (i.e., eigen value) were shown in the table below. We selected the top 19 (PC1-PC19) to participate in the outcome model development, based on the threshold of 1 for eigen values.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
| Standard deviation | 5.88098 | 3.20416 | 2.99159 | 2.64400 | 2.23198 | 1.98229 | 1.83906 | 1.73116 | 1.44335 |
| Proportion of Variance | 0.31159 | 0.09249 | 0.08063 | 0.06298 | 0.04488 | 0.0354 | 0.03047 | 0.02700 | 0.01877 |
| Cumulative Proportion | 0.31159 | 0.40408 | 0.4847 | 0.54768 | 0.59256 | 0.62796 | 0.65843 | 0.68543 | 0.70420 |
|  | PC10 | PC11 | PC12 | PC13 | PC14 | PC15 | PC16 | PC17 | PC18 |
| Standard deviation | 1.37762 | 1.30595 | 1.20848 | 1.15525 | 1.11346 | 1.08945 | 1.08093 | 1.06534 | 1.04166 |
| Proportion of Variance | 0.0171 | 0.01536 | 0.01316 | 0.01202 | 0.01117 | 0.01069 | 0.01053 | 0.01022 | 0.00978 |
| Cumulative Proportion | 0.7213 | 0.73666 | 0.74982 | 0.76184 | 0.77301 | 0.78371 | 0.79423 | 0.80446 | 0.81423 |
|  | PC19 | PC20 | PC21 | PC22 | PC23 | PC24 | PC25 |  |  |
| Standard deviation | 1.01626 | 0.97942 | 0.96837 | 0.94846 | 0.92998 | 0.91384 | 0.89119 |  |  |
| Proportion of Variance | 0.00930 | 0.00864 | 0.00845 | 0.00810 | 0.00779 | 0.00752 | 0.00716 |  |  |
| Cumulative Proportion | 0.82354 | 0.83218 | 0.84063 | 0.84873 | 0.85652 | 0.86405 | 0.87120 |  |  |

Proportion of Variance: The amount of variance the component accounts for in the data. For example, **PC1**accounts for **~31.2% of total variance**.

The following plot shows the contribution to the total variance by each individual principal component. The x-axis goes by the order of principal components (only top 50 are shown), and the y-axis is the contribution to the total variance by each individual component.

To understand the contributions of the original imaging features, we plotted the top (i.e., those with the largest absolute values) 50 coefficients that were used to transform the original features to the first principal component (PC1), which explained 31.2% of total variance. It is interesting to observe that for PC1, the features that contributed positively (especially the top 9) came from the first-order statistics feature category, while those that contributed negatively mainly came from the second order statistics feature category.



**CP patients:**

The top 25 with the highest standard deviation (i.e., eigen value) were shown in the table below. We selected the top 12 (PC1-PC12) to participate in the outcome model development, based on the threshold of 1 for eigen values.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 | PC8 | PC9 |
| Standard deviation | 5.61459 | 5.24775 | 3.87331 | 2.88656 | 2.50626 | 2.03962 | 1.66022 | 1.48960 | 1.42368 |
| Proportion of Variance | 0.284 | 0.2481 | 0.13516 | 0.07507 | 0.05659 | 0.03748 | 0.02483 | 0.01999 | 0.01826 |
| Cumulative Proportion | 0.284 | 0.53209 | 0.66725 | 0.74232 | 0.79891 | 0.83638 | 0.86122 | 0.88121 | 0.89947 |
|  | PC10 | PC11 | PC12 | PC13 | PC14 | PC15 | PC16 | PC17 | PC18 |
| Standard deviation | 1.37416 | 1.22902 | 1.07458 | 0.97167 | 0.94108 | 0.87178 | 0.73825 | 0.70350 | 0.65382 |
| Proportion of Variance | 0.01701 | 0.01361 | 0.0104 | 0.00851 | 0.00798 | 0.00685 | 0.00491 | 0.00446 | 0.00385 |
| Cumulative Proportion | 0.91648 | 0.93009 | 0.94049 | 0.94899 | 0.95697 | 0.96382 | 0.96873 | 0.97319 | 0.97704 |
|  | PC19 | PC20 | PC21 | PC22 | PC23 | PC24 | PC25 |  |  |
| Standard deviation | 0.61060 | 0.54998 | 0.51888 | 0.43758 | 0.43223 | 0.39842 | 0.37754 |  |  |
| Proportion of Variance | 0.00336 | 0.00273 | 0.00243 | 0.00172 | 0.00168 | 0.00143 | 0.00128 |  |  |
| Cumulative Proportion | 0.9804 | 0.98312 | 0.98555 | 0.98727 | 0.98896 | 0.99039 | 0.99167 |  |  |

Proportion of Variance: The amount of variance the component accounts for in the data. For example, **PC1**accounts for **~28.4% of total variance**.

The following plot shows the contribution to the total variance by each individual principal component. The x-axis goes by the order of principal components (only top 50 are shown), and the y-axis is the contribution to the total variance by each individual component.



To understand the contributions of the original imaging features, we plotted the top (i.e., those with the largest absolute values) 50 coefficients that were used to transform the original features to the first principal component (PC1), which explained 28.4% of total variance.

