Факультет компьютерных наук

Образовательная программа ПМИ

Topological data analysis of thoracic radiographic images shows improved radiomics-based lung tumor histology prediction

Подготовили студенты:

Пономарева Ольга
Теняев Александр
Алимханов Карим
Бабаев Минходж
Солодовников Михаил
Некрасов Артём
Хайрулин Инсаф

Рассматриваемые темы

1. Постановка задачи
2. Топологические признаки
3. Основные результаты
4. Сравнение моделей
5. Заключение

Постановка задачи

Цель: сравнение производительности моделей машинного обучения для предсказания гистологического типа опухоли на основе двух типов признаков: радиомических и топологических. Используются КТ снимки с контрастным веществом и без.

Постановка задачи

Исходные данные:

- X_{rad} - радиомические признаки, полученные из КТ-изображений (матрица признаков, где каждая строка - это наблюдение, а каждый столбец - признак);
- X_{top} - топологические признаки, полученные из тех же КТ-изображений;
- y - истинные метки классов (например, доброкачественная или злокачественная опухоль);
- C - бинарная переменная, показывающая наличие контрастного вещества.

Постановка задачи

Обучение моделей:

Для каждой комбинации признаков и условий (наличие контраста) обучались модели M_{rad}, M_{top}, M_{concat}, где

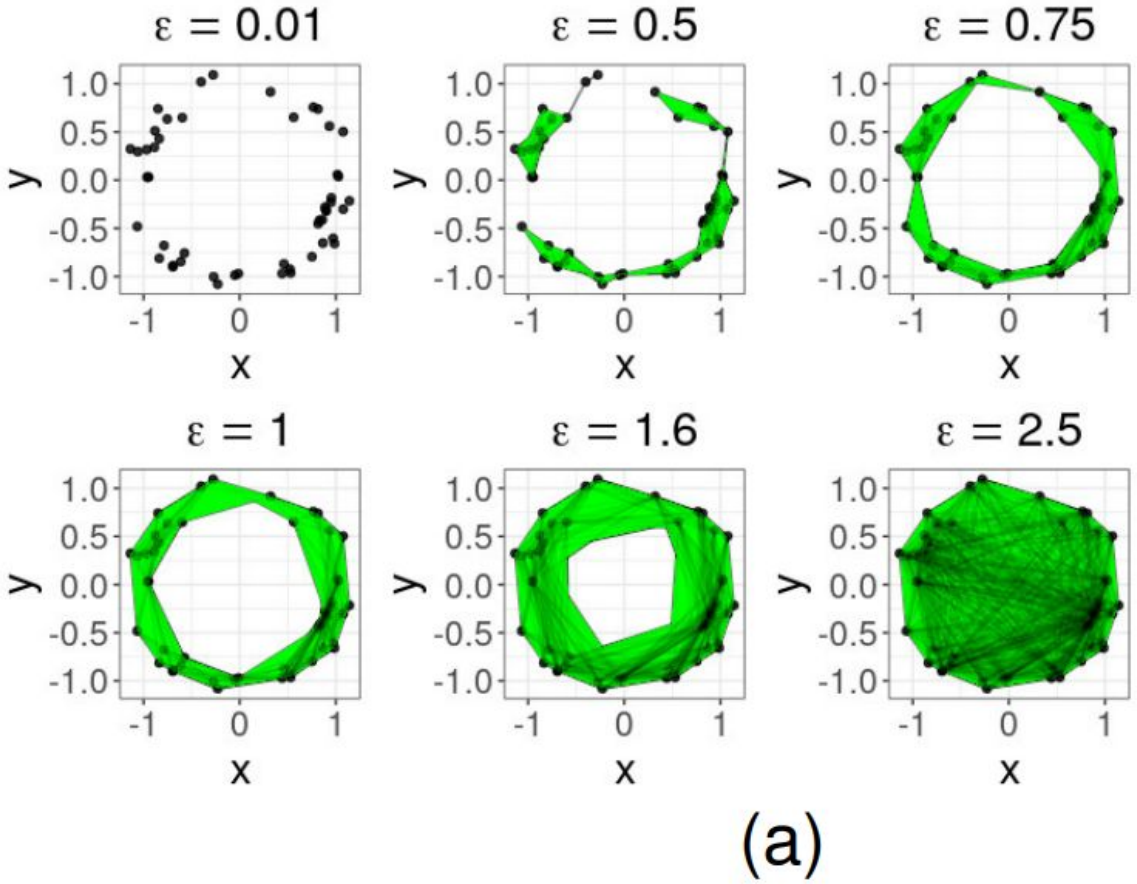
- M_{rad} - модель, обученная только на радиомических признаках X_{rad};
- M_{top} - модель, обученная только на топологических признаках X_{top};
- M_{concat} - модель, обученная на объединенных признаках $X_{rad} \cup X_{top}$;

Постановка задачи

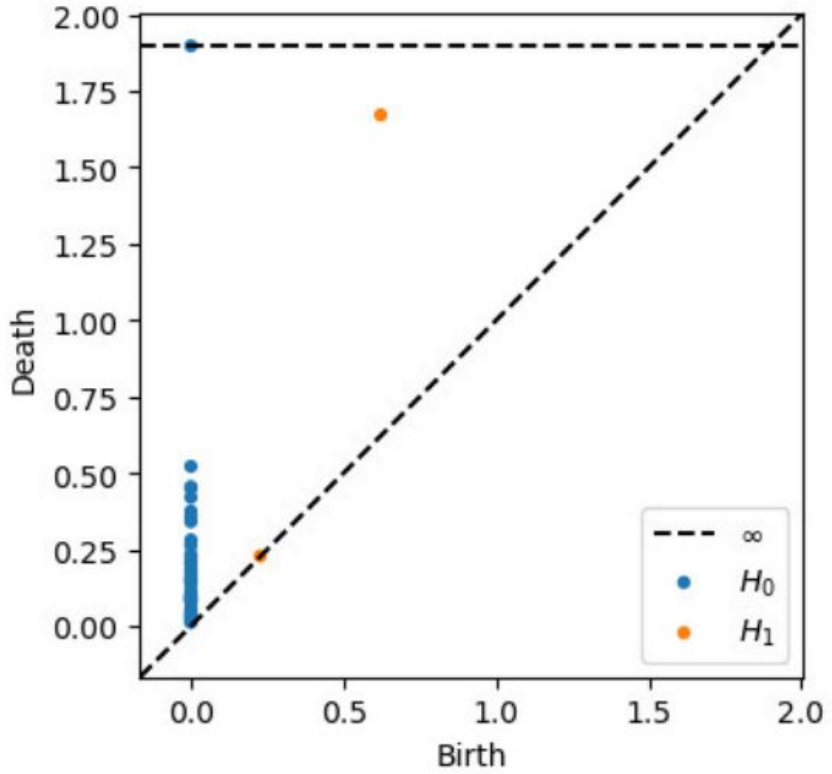
Пусть $f(x_i)$ - предсказание модели для объекта x_i, а y_i - истинная метка класса. Тогда производительность модели может быть измерена с помощью метрики ROC AUC (для классификации):

$$AUC(M) = \frac{1}{n}\sum_{i=1}^{n}[\mathbb{I}(f(x_i) > \text{threshold}, y_i = 1) + \mathbb{I}(f(x_i) \leq \text{threshold}, y_i = 0)]$$

Illustration of a filtration and holes in point clouds



(a)

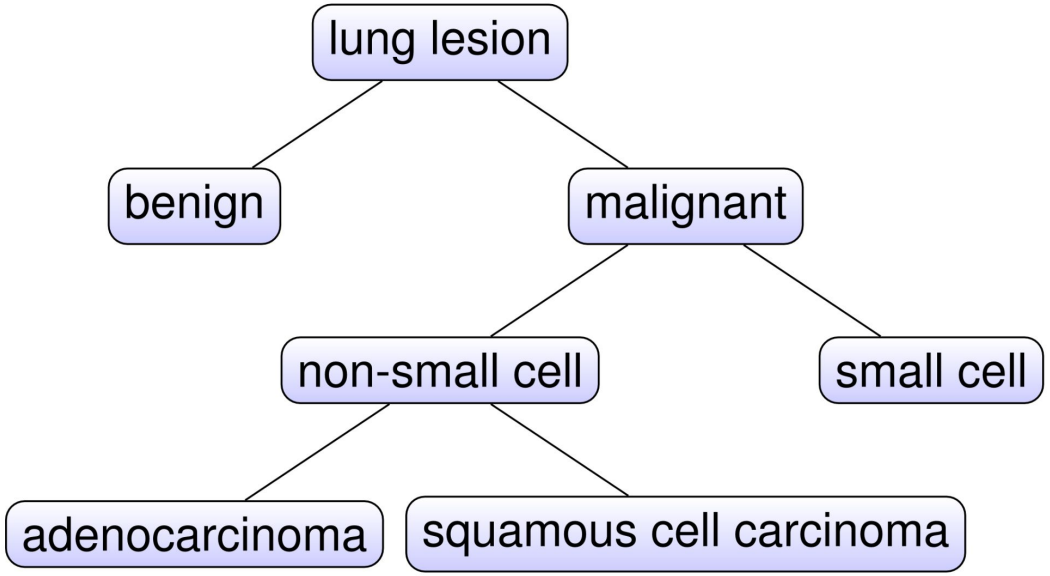


(b)

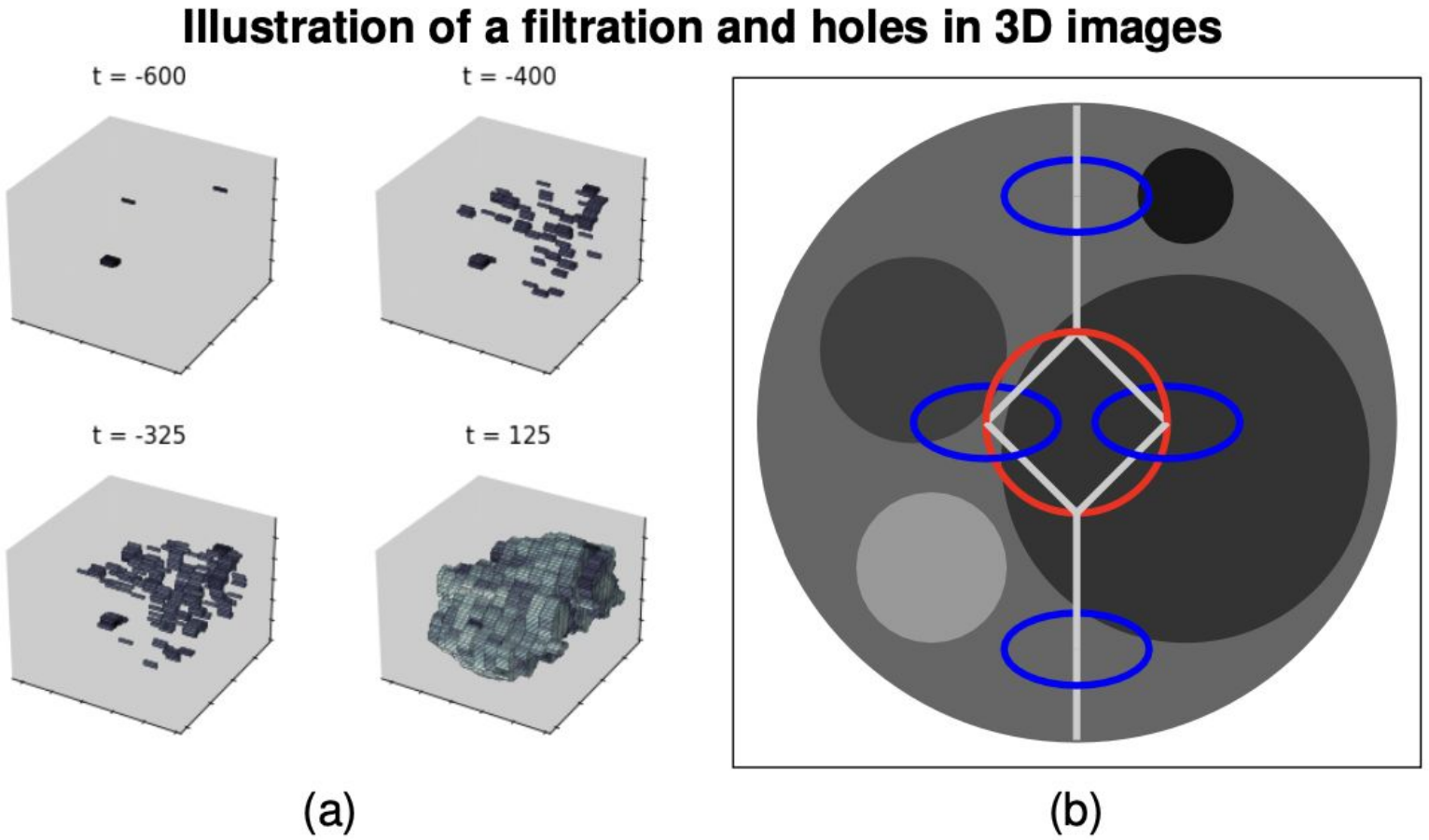
The number of observations for each class hierarchical structure of lung lesions

Table 1. The number of observations for each class of lung tumor in the data, with and without added contrast, in the San Francisco/Palo Alto (SF/PA) cohort and the Lung Image Database Consortium (LIDC)

	With contrast	Without contrast	Total
SF/PA			
benign	22	62	84
malignant	33	47	80
small	17	10	27
non-small	16	37	53
adeno	11	20	31
squamous	5	15	20
total	55	109	164
LIDC			
benign	24	5	29
malignant	17	8	25
total	41	13	54



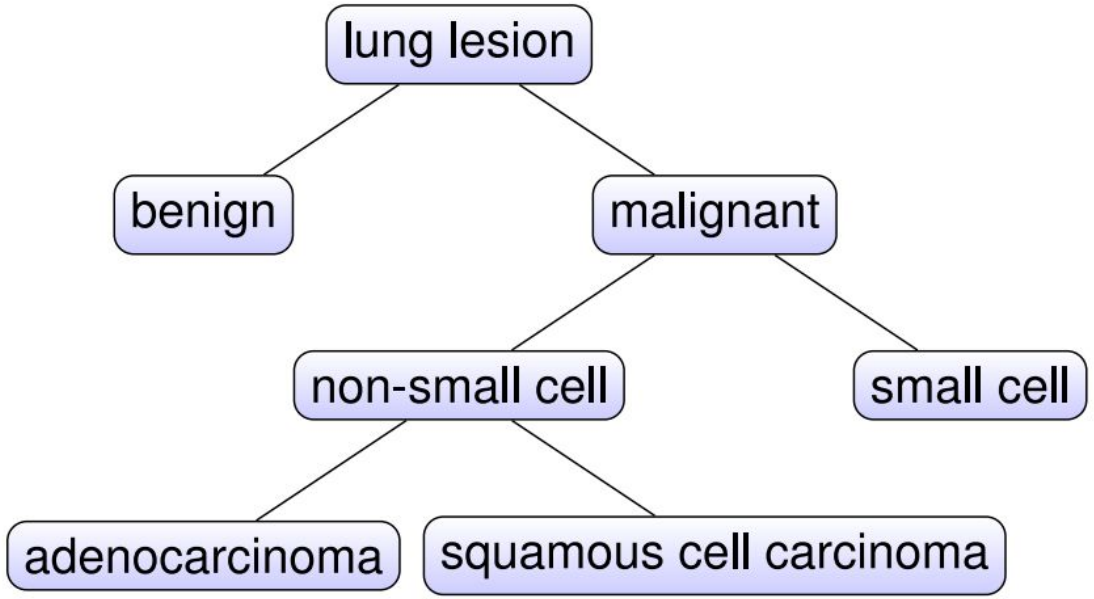
Quantitative image features extraction



Observation for each class and hierarchical structure of lung lesions

Table 1. The number of observations for each class of lung tumor in the data, with and without added contrast, in the San Francisco/Palo Alto (SF/PA) cohort and the Lung Image Database Consortium (LIDC)

	With contrast	Without contrast	Total
SF/PA			
benign	22	62	84
malignant	33	47	80
small	17	10	27
non-small	16	37	53
adeno	11	20	31
squamous	5	15	20
total	55	109	164
LIDC			
benign	24	5	29
malignant	17	8	25
total	41	13	54



Радиомические признаки

Изображения и маски пересэмплированы до размеров 1x1x1 mm^3

С помощью PyRadiomics выявлены 105 признаков, которые относятся к следующим категориям:

1. Статистики первого порядка
2. Признаки на основе формы в 3D
3. Матрицы зон размерности градации серостей (GLSZM)
4. Матрицы совместной встречаемости градации серостей (GLCM)
5. Матрицы длины пробега градации серостей (GLRLM)
6. Матрицы разницы тонов соседних серостей (NGTDM)
7. Матрицы зависимости градации серостей

Топологические признаки

Из каждого скана были извлечены различные типы диаграмм персистентности.

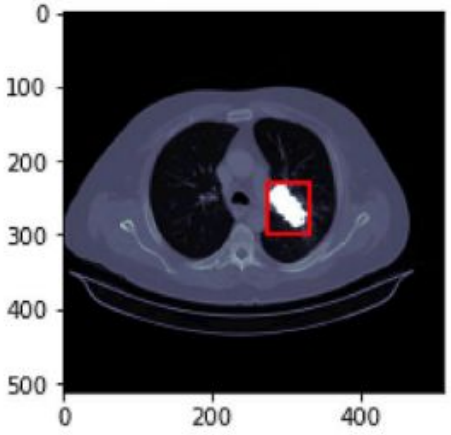
Работа диаграммы персистентности:

1. Обработка данных
2. Создание симплициальных комплексов
3. Оценка топологических “дыр”
4. Пары “рождения - смерти”

Размерности топологических “дыр”:

- Размерность 0: Связные компоненты. *Это отдельные участки опухоли без выходов к другим частям.*
- Размерность 1: Циклы. *Это замкнутые пути на поверхности опухоли.*
- Размерность 2: Пустоты. *Это внутренние пространства в опухоли.*

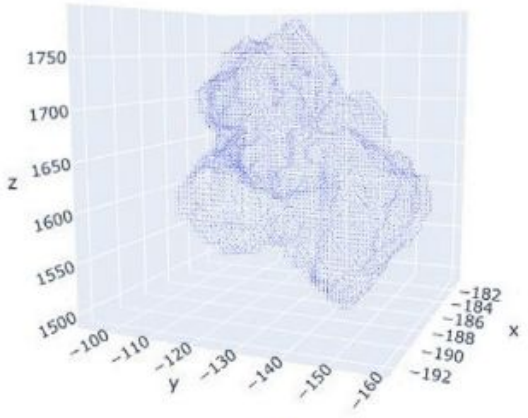
Графики



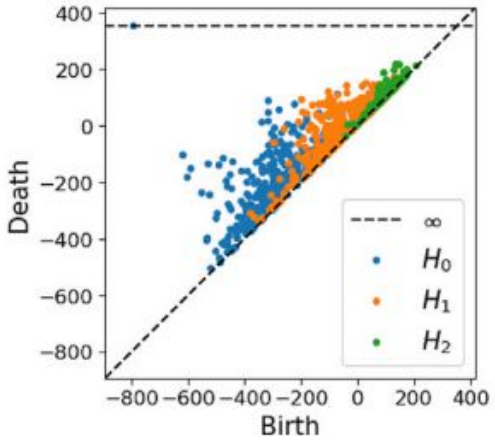
A

B

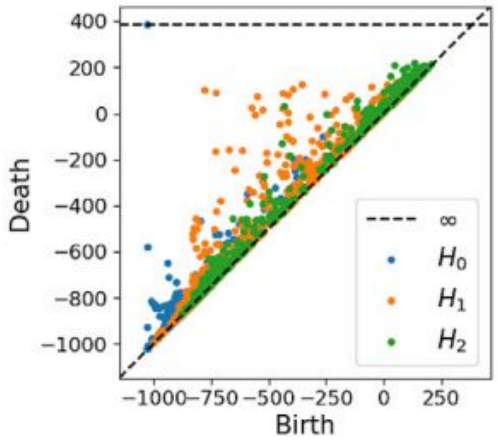
C



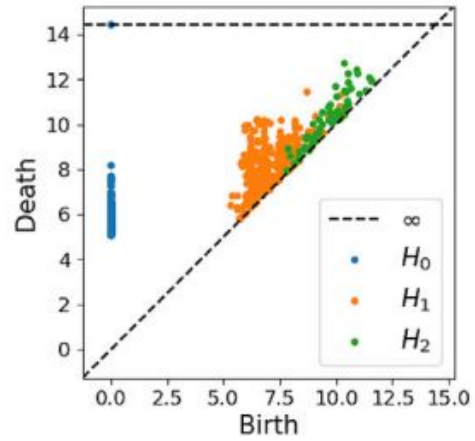
D



E



F



G

Supervised ML modeling

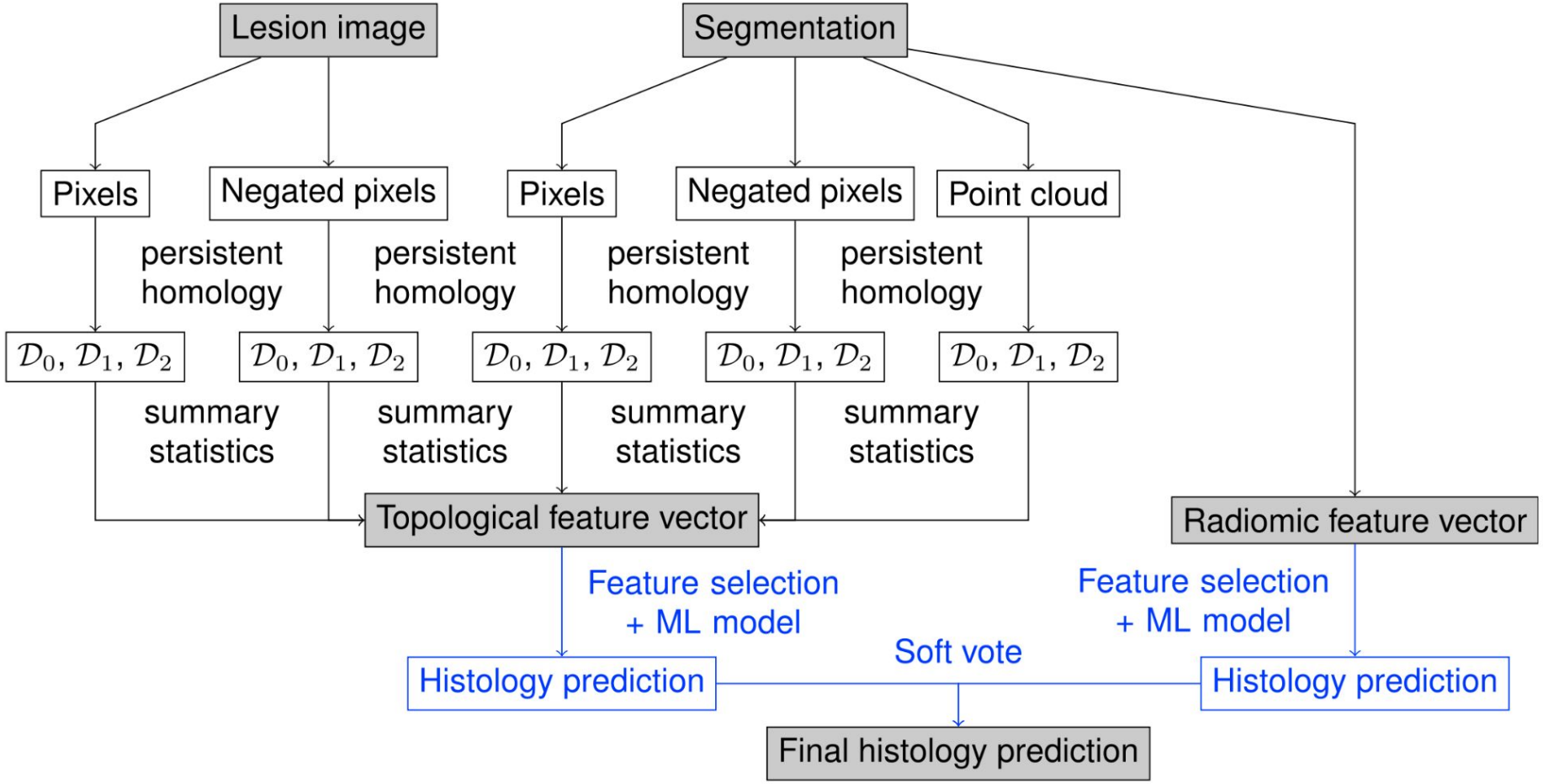
Feature preprocessing:

- mean imputation
- min-max normalization
- discretization
- mRMR-method

Models:

- logistic/linear regression (LR)
- random forest classification/regression (RF)
- k-nearest neighbor classification/regression (KNN)
- support vector machine/regressor (SV)
- Gaussian naive Bayes classification/Bayesian regression (BAY)
- extreme gradient-boosted trees classification/regression(XGB)

Pipeline



Pipeline used to evaluate and compare topological and radiomics features to predict the histology of lung tumors

Model evaluation

- 10 repeats of 5-fold cross-validation
- AUC демонстрирует, насколько хорошо модель может различать, например, доброкачественные и злокачественные опухоли
- R2 показывает, насколько точно предсказывает модель регрессии
- Rad,Top,Concat - радиомические, топологические признаки и их объединение
- Stack, Vote - ансамблевые модели

Lung tumor histology prediction benign vs malignant

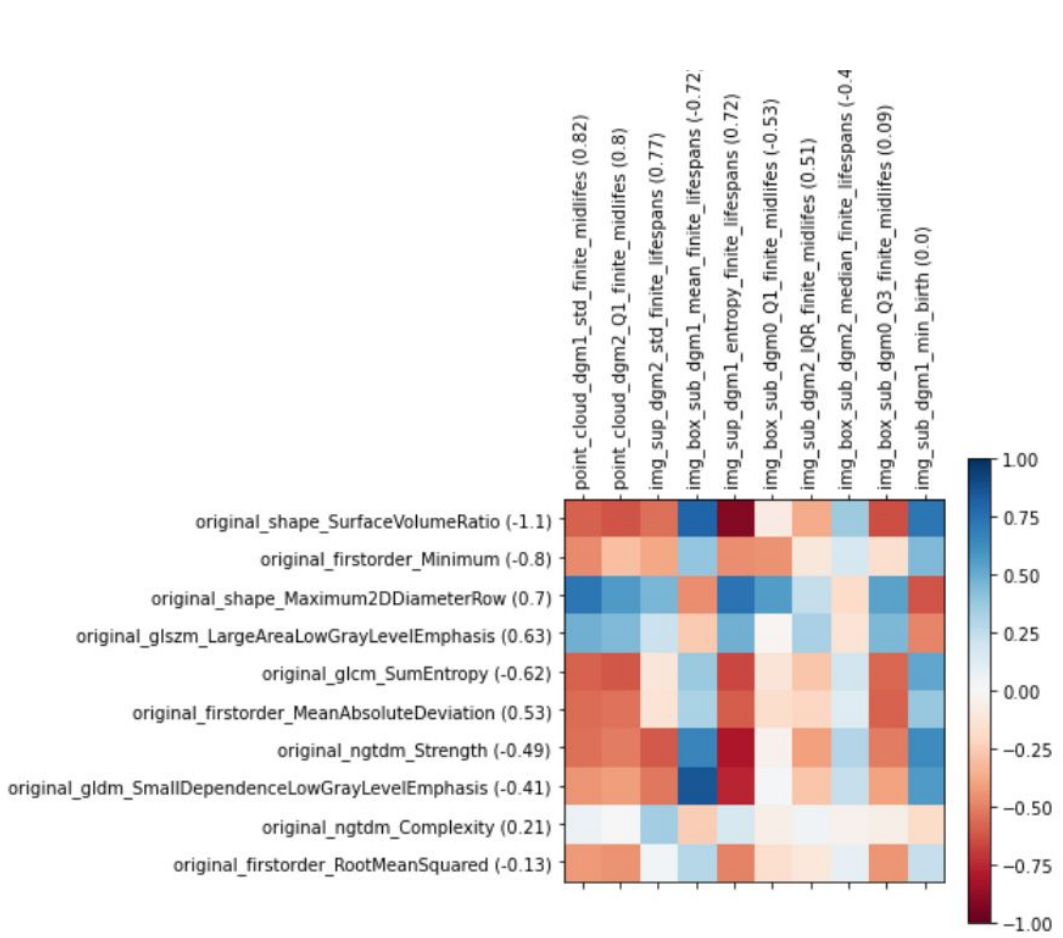
Table 2. Mean performances in percentage (ROC AUC for classification and r^2 for regression) for lung tumor histology prediction

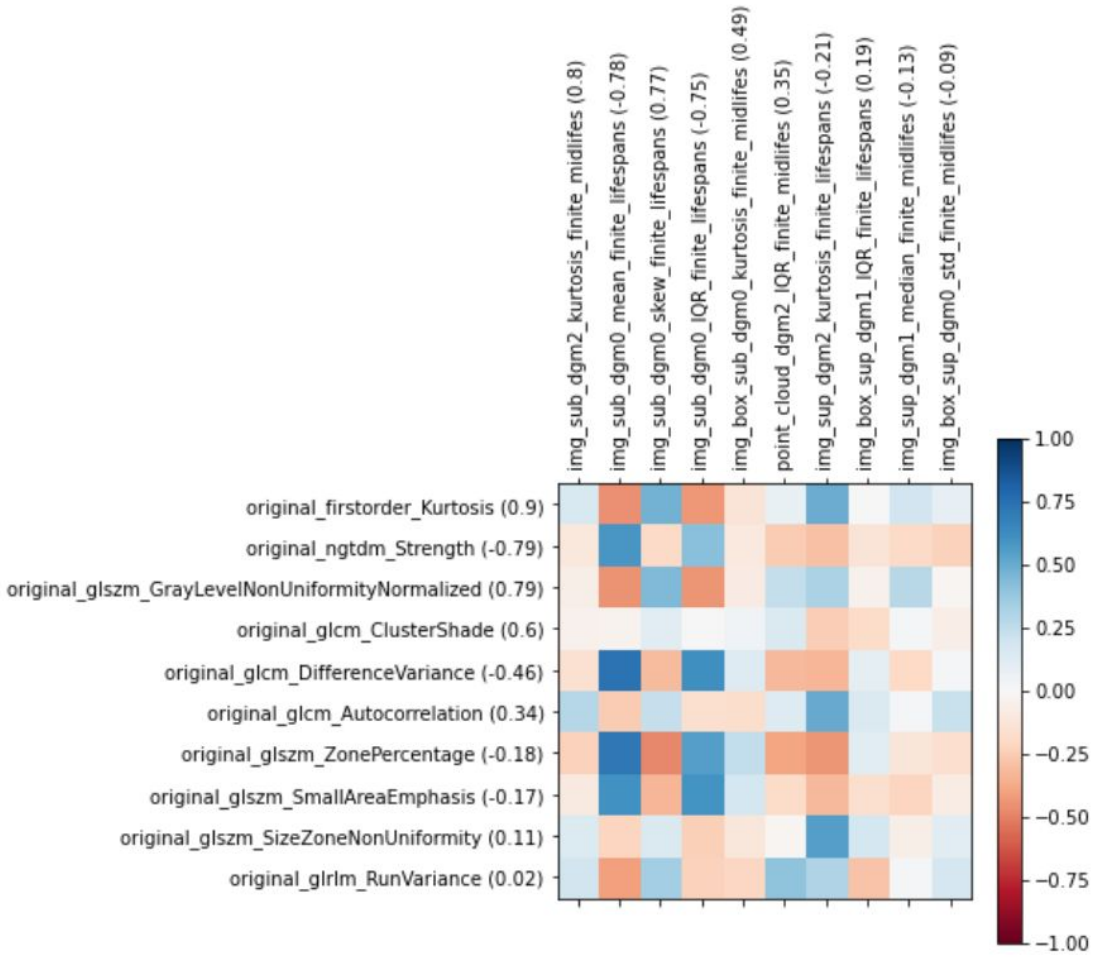
Problem	C	SEM	Rad	Top	Concat	Vote	Stack	Best model	Best score	p vote ≥ rad
Benign versus	Y	–	84.6	86.8	86.7	87.9[a]	85.8	LR + vote	88.9	$5.7 \cdot 10^{-5}$
malignant (SF/PA)	N	–	74.0	75.7	76.5	78.2[a]	73.8	LR + vote	80.2	$1.7 \cdot 10^{-7}$
Small cell versus	Y	–	77.5[a]	62.7	66.1	75.0	71.7	LR + rad only	79.8	0.94
non-small cell (SF/PA)	N	–	80.6	78.6	80.9	83.4[a]	75.9	RF + vote	86.8	$3.9 \cdot 10^{-2}$
Adeno versus	Y	–	67.2	91.2[a]	90.1	88.3	–	RF + top/concat	98.3	$1.2 \cdot 10^{-17}$
squamous (SF/PA)	N	–	64.3	70.0	68.8	71.2[a]	65.1	BAY + vote	75.0	$3.8 \cdot 10^{-5}$
Malignancy	Y	61.1	56.3	52.0	53.4	59.0[a]	53.5	RF + vote	61.3	$5.6 \cdot 10^{-7}$
regression (LIDC)	N	54.2	42.8	36.4	38.2	45.8[a]	38.8	RF + vote	49.0	$3.3 \cdot 10^{-9}$
Benign versus	Y	66.9	58.2	61.6[a]	59.3	60.1	56.6	KNN + stack	67.7	0.11
malignant (LIDC)	N	15.6	54.1	63.1	66.2[a]	61.5	43.3	XGB + vote	78.0	$1.6 \cdot 10^{-2}$

C, whether contrast material was added (Y) or not (N); SEM, semantic features that were manually assigned by expert radiologists; rad, radiomic features; top, topological features; concat, concatenated radiomic and topological features; vote, voting ensemble; stack, stacking ensemble; p vote ≥ rad, p value for the null hypothesis that the mean performance when using solely radiomic features is at least as good as using both radiomic and topological features through a voting ensemble.

[a]Best mean performances with automated features.

Feature correlation for benign vs. malignant (classification, SF/PA, with/without contrast)



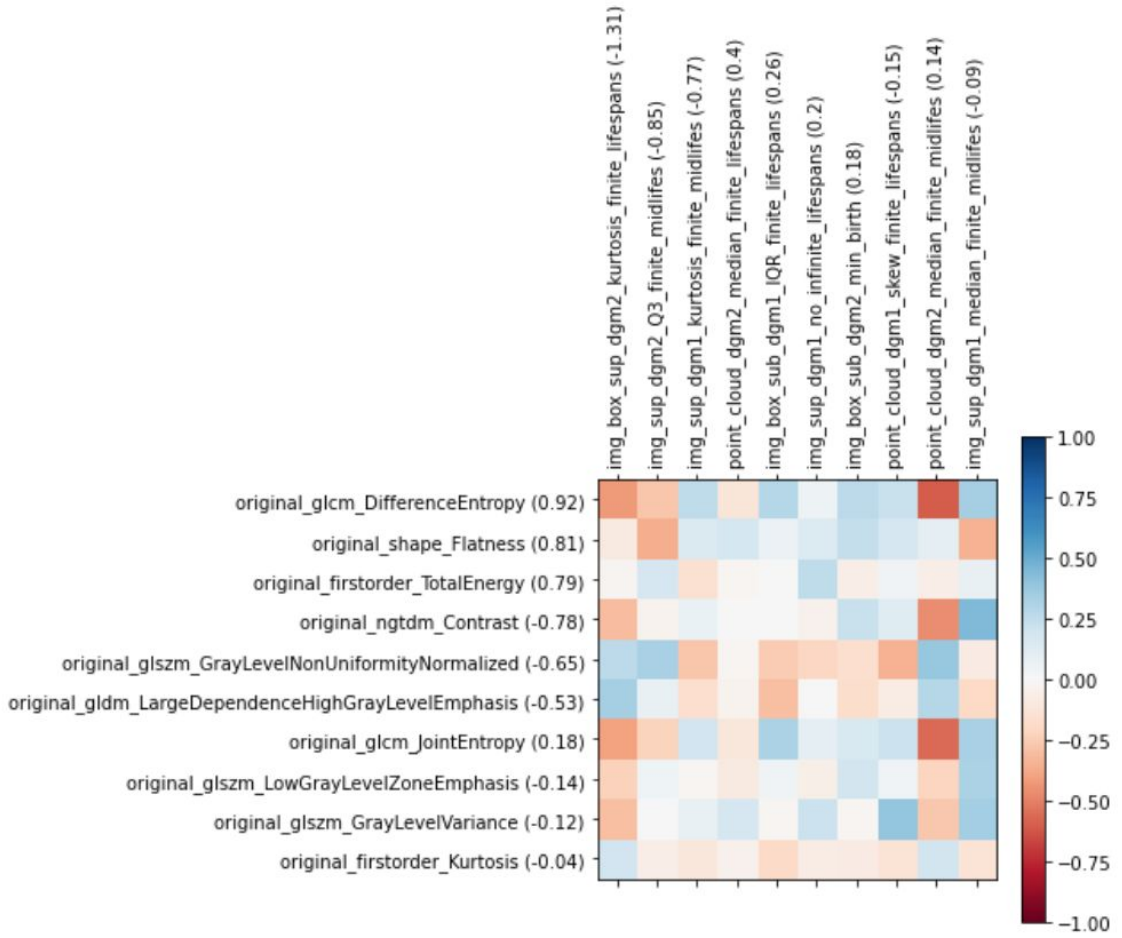


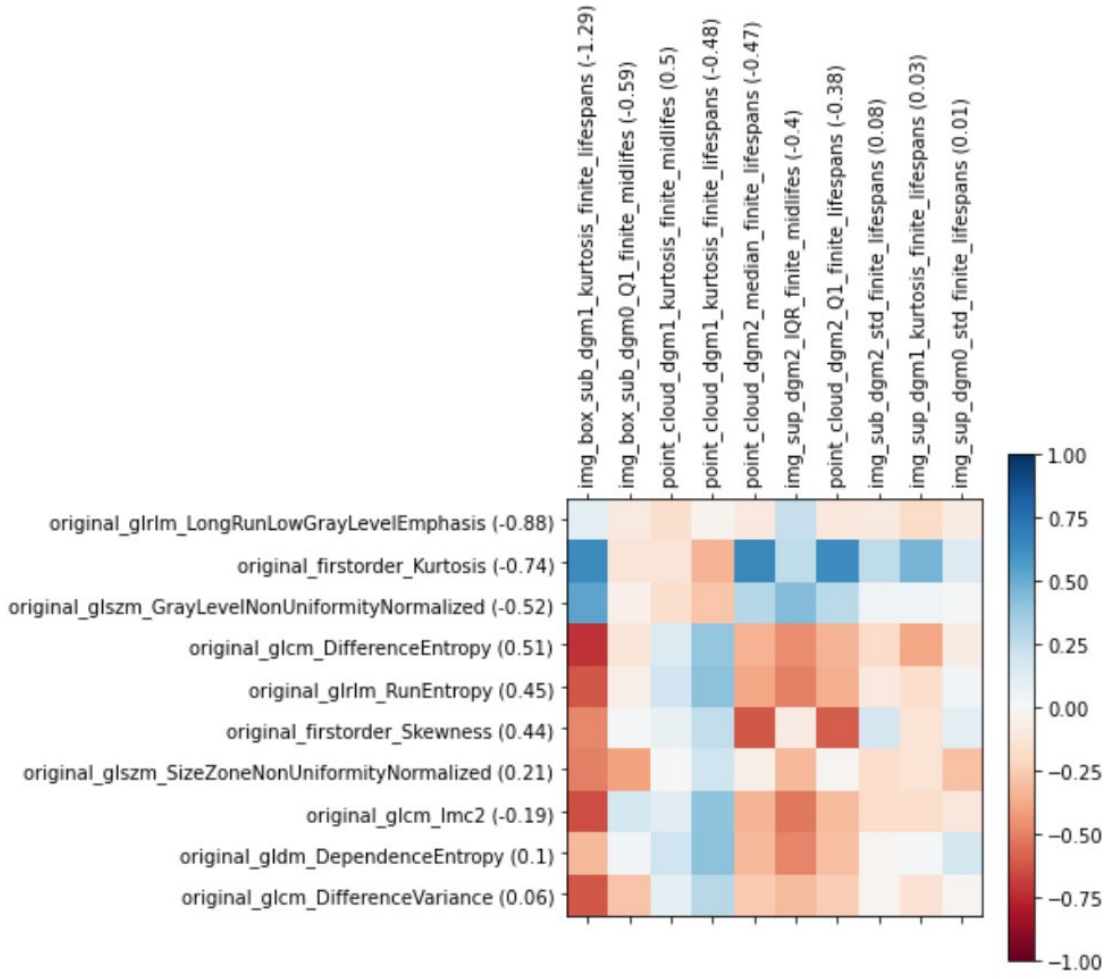
Lung tumor histology prediction small cell vs non-small

Table 2. Mean performances in percentage (ROC AUC for classification and r^2 for regression) for lung tumor histology prediction

Problem	C	SEM	Rad	Top	Concat	Vote	Stack	Best model	Best score	p vote ≥ rad
Benign versus	Y	–	84.6	86.8	86.7	87.9[a]	85.8	LR + vote	88.9	$5.7 \cdot 10^{-5}$
malignant (SF/PA)	N	–	74.0	75.7	76.5	78.2[a]	73.8	LR + vote	80.2	$1.7 \cdot 10^{-7}$
Small cell versus	Y	–	77.5[a]	62.7	66.1	75.0	71.7	LR + rad only	79.8	0.94
non-small cell (SF/PA)	N	–	80.6	78.6	80.9	83.4[a]	75.9	RF + vote	86.8	$3.9 \cdot 10^{-2}$
Adeno versus	Y	–	67.2	91.2[a]	90.1	88.3	–	RF + top/concat	98.3	$1.2 \cdot 10^{-17}$
squamous (SF/PA)	N	–	64.3	70.0	68.8	71.2[a]	65.1	BAY + vote	75.0	$3.8 \cdot 10^{-5}$
Malignancy	Y	61.1	56.3	52.0	53.4	59.0[a]	53.5	RF + vote	61.3	$5.6 \cdot 10^{-7}$
regression (LIDC)	N	54.2	42.8	36.4	38.2	45.8[a]	38.8	RF + vote	49.0	$3.3 \cdot 10^{-9}$
Benign versus	Y	66.9	58.2	61.6[a]	59.3	60.1	56.6	KNN + stack	67.7	0.11
malignant (LIDC)	N	15.6	54.1	63.1	66.2[a]	61.5	43.3	XGB + vote	78.0	$1.6 \cdot 10^{-2}$

Feature correlation for small-cell vs. non-small cell (classification, SF/PA, with/without contrast)





Lung tumor histology prediction squamous vs ADC

Table 2. Mean performances in percentage (ROC AUC for classification and r^2 for regression) for lung tumor histology prediction

Problem	C	SEM	Rad	Top	Concat	Vote	Stack	Best model	Best score	p vote ≥ rad
Benign versus	Y	–	84.6	86.8	86.7	87.9[a]	85.8	LR + vote	88.9	$5.7 \cdot 10^{-5}$
malignant (SF/PA)	N	–	74.0	75.7	76.5	78.2[a]	73.8	LR + vote	80.2	$1.7 \cdot 10^{-7}$
Small cell versus	Y	–	77.5[a]	62.7	66.1	75.0	71.7	LR + rad only	79.8	0.94
non-small cell (SF/PA)	N	–	80.6	78.6	80.9	83.4[a]	75.9	RF + vote	86.8	$3.9 \cdot 10^{-2}$
Adeno versus	Y	–	67.2	91.2[a]	90.1	88.3	–	RF + top/concat	98.3	$1.2 \cdot 10^{-17}$
squamous (SF/PA)	N	–	64.3	70.0	68.8	71.2[a]	65.1	BAY + vote	75.0	$3.8 \cdot 10^{-5}$
Malignancy	Y	61.1	56.3	52.0	53.4	59.0[a]	53.5	RF + vote	61.3	$5.6 \cdot 10^{-7}$
regression (LIDC)	N	54.2	42.8	36.4	38.2	45.8[a]	38.8	RF + vote	49.0	$3.3 \cdot 10^{-9}$
Benign versus	Y	66.9	58.2	61.6[a]	59.3	60.1	56.6	KNN + stack	67.7	0.11
malignant (LIDC)	N	15.6	54.1	63.1	66.2[a]	61.5	43.3	XGB + vote	78.0	$1.6 \cdot 10^{-2}$

LIDC

Performances for malignancy prediction (regression, LIDC, with contrast)

model	sem	rad	top	concat	vote	stack
LR	61.3 ± 6.4	57.5 ± 6.1	53.3 ± 5.6	53.3 ± 5.5	58.5 ± 5.5	$\mathbf{58.6 \pm 5.7}$
RF	65.3 ± 6.4	57.3 ± 7.0	54.1 ± 7.0	55.5 ± 6.3	$\mathbf{61.3 \pm 6.1}$	52.6 ± 6.8
KNN	61.6 ± 6.9	57.5 ± 6.4	48.7 ± 10.1	51.8 ± 8.7	$\mathbf{59.4 \pm 6.5}$	51.9 ± 7.3
SV	59.9 ± 7.5	56.4 ± 7.1	51.1 ± 6.0	52.4 ± 5.8	57.5 ± 6.1	$\mathbf{57.6 \pm 6.5}$
BAY	61.4 ± 6.3	57.6 ± 5.9	53.5 ± 5.6	53.6 ± 5.4	58.5 ± 5.5	$\mathbf{58.6 \pm 5.8}$
XGB	57.0 ± 7.3	51.3 ± 9.2	51.4 ± 8.0	53.5 ± 7.1	$\mathbf{59.0 \pm 7.1}$	41.9 ± 8.4
mean	61.1 ± 7.2	56.3 ± 7.4	52.0 ± 7.5	53.4 ± 6.7	$\mathbf{59.0 \pm 6.3}$	53.5 ± 9.0

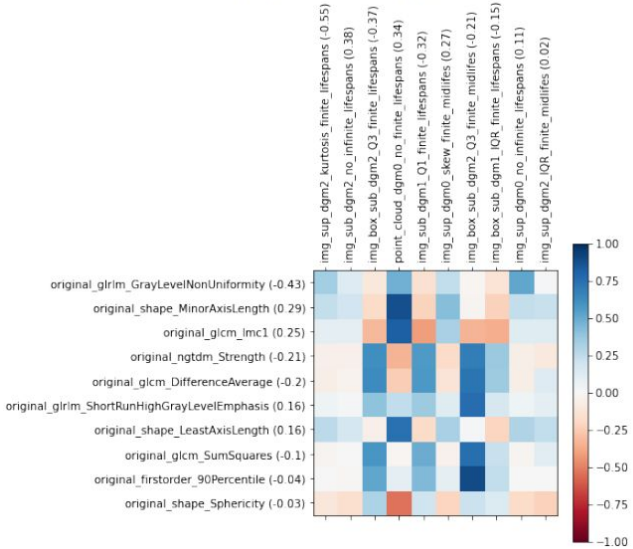
TABLE S7. r^2 performances in % with standard deviations for continuous *malignancy* outcome prediction of lung tumor nodules from CT scan images *with added contrast*, using semantic features (*sem*), radiomic features (*rad*) and topological features (*top*), as well as for three models combining both: through concatenation (*concat*), soft voting (*vote*), and stacking. Each scores is averaged over 50 models, obtained through 10-repeated samplings in 5 folds. Non-semantic best scores are marked in bold.

Performances for malignancy prediction (regression, LIDC, without contrast)

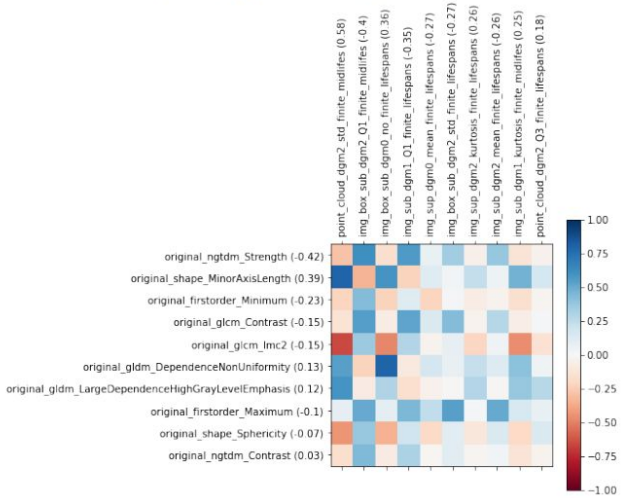
model	sem	rad	top	concat	vote	stack
LR	54.8 ± 4.9	43.3 ± 5.5	35.6 ± 7.1	36.4 ± 6.9	44.2 ± 5.2	$\mathbf{45.1 \pm 5.3}$
RF	56.9 ± 5.1	45.3 ± 6.0	41.0 ± 8.0	43.6 ± 6.4	$\mathbf{49.0 \pm 5.6}$	36.4 ± 7.9
KNN	54.2 ± 6.1	39.6 ± 7.3	31.8 ± 9.9	35.4 ± 9.4	$\mathbf{45.2 \pm 6.2}$	34.5 ± 8.2
SV	54.1 ± 5.1	42.1 ± 5.8	34.8 ± 7.2	35.5 ± 7.3	43.8 ± 5.4	$\mathbf{44.3 \pm 5.6}$
BAY	54.8 ± 4.9	43.5 ± 5.4	35.7 ± 7.0	36.6 ± 6.9	44.1 ± 5.2	$\mathbf{45.1 \pm 5.3}$
XGB	50.6 ± 5.3	43.0 ± 6.4	39.6 ± 8.1	41.7 ± 6.6	$\mathbf{48.3 \pm 5.7}$	27.4 ± 11.4
mean	54.2 ± 5.6	42.8 ± 6.3	36.4 ± 8.5	38.2 ± 8.0	$\mathbf{45.8 \pm 6.0}$	38.9 ± 10.1

TABLE S8. r^2 performances in % with standard deviations for continuous *malignancy* outcome prediction of lung tumor nodules from CT scan images *without added contrast*, using semantic features (*sem*), radiomic features (*rad*) and topological features (*top*), as well as for three models combining both: through concatenation (*concat*), soft voting (*vote*), and stacking. Each scores is averaged over 50 models, obtained through 10-repeated samplings in 5 folds. Non-semantic best scores are marked in bold.

Feature correlation for malignancy prediction (regression, LIDC, with contrast)



Feature correlation for malignancy prediction (regression, LIDC, without contrast)



LIDC

Performances for benign vs. malignant (classification, LIDC, with contrast)

model	sem	rad	top	concat	vote	stack
LR	66.3 ± 20.2	53.6 ± 19.3	$\mathbf{60.3 \pm 18.9}$	56.4 ± 17.5	57.5 ± 20.2	54.2 ± 20.7
RF	68.2 ± 19.3	57.0 ± 18.4	$\mathbf{61.0 \pm 20.2}$	59.3 ± 19.4	60.6 ± 18.0	50.0 ± 22.1
KNN	66.8 ± 20.1	66.3 ± 17.8	61.7 ± 19.0	62.4 ± 18.6	64.5 ± 18.7	$\mathbf{67.7 \pm 16.4}$
SV	67.7 ± 20.8	56.7 ± 20.6	$\mathbf{59.3 \pm 19.8}$	57.0 ± 15.0	53.9 ± 20.7	52.8 ± 19.7
BAY	64.3 ± 19.5	57.4 ± 20.3	$\mathbf{61.9 \pm 19.0}$	59.7 ± 18.8	61.3 ± 19.4	55.3 ± 22.3
XGB	68.0 ± 16.3	57.9 ± 17.6	$\mathbf{65.7 \pm 17.4}$	61.1 ± 17.8	63.0 ± 17.5	59.5 ± 21.5
mean	66.9 ± 19.5	58.2 ± 19.4	$\mathbf{61.6 \pm 19.2}$	59.3 ± 18.0	60.1 ± 19.5	56.6 ± 21.3

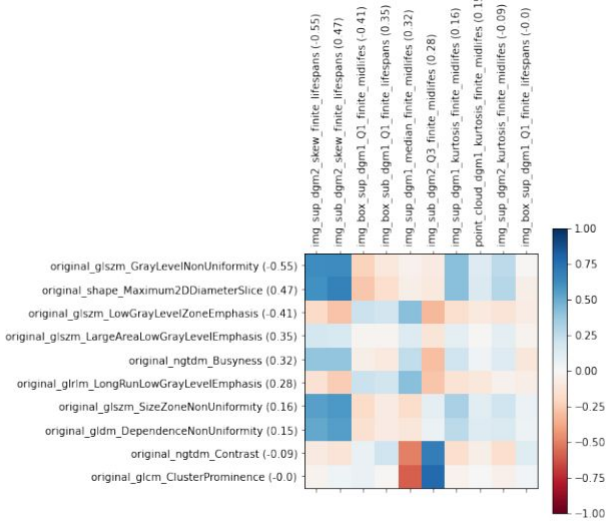
TABLE S9. ROC AUC performances in % with standard deviations for *benign vs. malignant* classification of lung tumor nodules from CT scan images *with added contrast*, using semantic features (*sem*), radiomic features (*rad*) and topological features (*top*), as well as for three models combining both: through concatenation (*concat*), soft voting (*vote*), and stacking. Each scores is averaged over 50 models, obtained through 10-repeated samplings in 5 folds. Non-semantic best scores are marked in bold.

Performances for benign vs. malignant (classification, LIDC, without contrast)

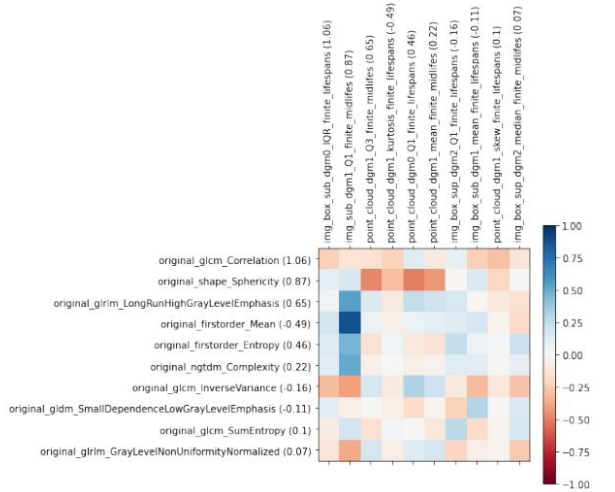
model	sem	rad	top	concat	vote	stack
LR	16.0 ± 35.3	58.3 ± 46.2	62.0 ± 43.1	$\mathbf{70.3 \pm 42.3}$	53.0 ± 47.7	43.7 ± 43.2
RF	9.3 ± 21.4	57.3 ± 42.6	59.7 ± 41.4	$\mathbf{65.3 \pm 42.7}$	60.0 ± 42.6	24.7 ± 37.0
KNN	29.3 ± 34.6	45.8 ± 33.9	58.0 ± 36.9	$\mathbf{59.5 \pm 37.7}$	52.7 ± 44.7	45.0 ± 32.5
SV	12.0 ± 30.9	$\mathbf{64.3 \pm 45.8}$	56.0 ± 43.2	63.7 ± 43.8	49.0 ± 45.3	51.3 ± 41.3
BAY	14.8 ± 26.2	57.7 ± 40.3	66.5 ± 32.3	64.5 ± 30.5	$\mathbf{76.2 \pm 35.1}$	46.2 ± 43.1
XGB	12.3 ± 21.6	40.8 ± 33.2	76.5 ± 34.8	74.0 ± 38.4	$\mathbf{78.0 \pm 34.9}$	49.0 ± 36.7
mean	15.6 ± 29.6	54.1 ± 41.5	63.1 ± 39.4	$\mathbf{66.2 \pm 39.8}$	61.5 ± 43.5	43.3 ± 40.1

TABLE S10. ROC AUC performances in % with standard deviations for *benign vs. malignant* classification of lung tumor nodules from CT scan images *without added contrast*, using semantic features (*sem*), radiomic features (*rad*) and topological features (*top*), as well as for three models combining both: through concatenation (*concat*), soft voting (*vote*), and stacking. Each scores is averaged over 50 models, obtained through 10-repeated samplings in 5 folds. Non-semantic best scores are marked in bold.

Feature correlation for benign vs. malignant (classification, LIDC, with contrast)



Feature correlation for benign vs. malignant (classification, LIDC, without contrast)



Results

Table 2. Mean performances in percentage (ROC AUC for classification and r^2 for regression) for lung tumor histology prediction

Problem	C	SEM	Rad	Top	Concat	Vote	Stack	Best model	Best score	p vote ≥ rad
Benign versus	Y	–	84.6	86.8	86.7	87.9[a]	85.8	LR + vote	88.9	$5.7 \cdot 10^{-5}$
malignant (SF/PA)	N	–	74.0	75.7	76.5	78.2[a]	73.8	LR + vote	80.2	$1.7 \cdot 10^{-7}$
Small cell versus	Y	–	77.5[a]	62.7	66.1	75.0	71.7	LR + rad only	79.8	0.94
non-small cell (SF/PA)	N	–	80.6	78.6	80.9	83.4[a]	75.9	RF + vote	86.8	$3.9 \cdot 10^{-2}$
Adeno versus	Y	–	67.2	91.2[a]	90.1	88.3	–	RF + top/concat	98.3	$1.2 \cdot 10^{-17}$
squamous (SF/PA)	N	–	64.3	70.0	68.8	71.2[a]	65.1	BAY + vote	75.0	$3.8 \cdot 10^{-5}$
Malignancy	Y	61.1	56.3	52.0	53.4	59.0[a]	53.5	RF + vote	61.3	$5.6 \cdot 10^{-7}$
regression (LIDC)	N	54.2	42.8	36.4	38.2	45.8[a]	38.8	RF + vote	49.0	$3.3 \cdot 10^{-9}$
Benign versus	Y	66.9	58.2	61.6[a]	59.3	60.1	56.6	KNN + stack	67.7	0.11
malignant (LIDC)	N	15.6	54.1	63.1	66.2[a]	61.5	43.3	XGB + vote	78.0	$1.6 \cdot 10^{-2}$

C, whether contrast material was added (Y) or not (N); SEM, semantic features that were manually assigned by expert radiologists; rad, radiomic features; top, topological features; concat, concatenated radiomic and topological features; vote, voting ensemble; stack, stacking ensemble; p vote ≥ rad, p value for the null hypothesis that the mean performance when using solely radiomic features is at least as good as using both radiomic and topological features through a voting ensemble.

[a]Best mean performances with automated features.

Основные выводы

1. Топологический анализ данных улучшает точность предсказания гистологии опухолей легких
2. Топологические признаки превосходят радиомические для предсказания истинной гистологии опухолей
3. Топологический анализ менее полезен для классификации мелкоклеточных и немелкоклеточных опухолей
4. Использование контрастного материала в изображениях улучшает точность предсказания
5. Комбинация радиомических и топологических признаков даёт лучшие результаты
6. Перспективы дальнейших исследований и применения TDA