Klasifikácia Snímok Tkaniva Hrubého čreva s Využitím Analýzy Textúrnych Deskriptorov a Strojového Učenia

Oliver Sidor

6. januára 2026

1 Úvod

Detekcia a klasifikácia nádorových buniek na snímkach tkaniva hrubého čreva je rozsiahla oblasť patologickej medicíny. Kolorektálny karcinóm je jedným z najčastejších typov rakoviny, ktorý si každoročne vyžiada tisícky ľudských životov.

Hlavným problémom je subjektivita rozhodovania doktorov medzi sebou a taktiež variabilita rozhodnutí konkrétneho doktora pri detekovaní a určovaní úrovne nádoru.

Počítačové videnie a strojové učenie prinášajú revolučné možnosti pre automatizovanú a objektívnu detekciu nádorov v medicínskych snímkach. Tieto metódy môžu slúžiť ako:

1. Podporný nástroj pre lekárov pri diagnostike
2. Screeningový systém tkanív
3. Referencia pre vzdelávanie a štandardizáciu
4. Podporné merítko kvality vyšetrenia

1.1 Ciele Práce

Hlavné ciele tejto práce sú:

1. Implementovať kompletnú pipeline na diagnostiku tkaniva
2. Porovnať výkon viacerých algoritmov strojového učenia na binárnej klasifikácií
3. Zaistiť reprodukovateľnosť výsledkov

2 Datasety

2.0.1 LC25000 Dataset

Primárnym zdrojom dát bol verejný LC25000 dataset (Lung and Colon Cancer Histopathological Images), ktorý je voľne dostupný pre výskumné účely. Dataset obsahuje:

Charakteristika	**Hodnota**
Celkový počet vzoriek	10 000 snímok
Zdravé snímky (colon_n)	5 000 snímok
Snímky obsahujúce nádor (colon_aca)	5 000 snímok
Rozlíšenie snímky	768 × 768 pixelov
Formát	JPEG farebné snímky
Typ tkaniva	Tkanivo hrubého čreva

Tabuľka 1: Špecifikácia LC25000 datasetu

2.0.2 Dataset CRC-HGD-v1

Ďalším zdrojom dát bol verejný CRC-HGD-v1 dataset (A Histopathological Image Dataset for Grading Colorectal Cancer - Mendeley Data), ktorý je taktiež voľne dostupný pre výskumné účely. Dataset obsahuje:

Charakteristika	**Hodnota**
Celkový počet vzoriek	2304 snímok
Zdravé snímky (colon_n)	8 snímok
Snímky obsahujúce nádor - úroveň 1	424 snímok
Snímky obsahujúce nádor - úroveň 2	300 snímok
Snímky obsahujúce nádor - úroveň 3	132 snímok
Rozlíšenie snímky	800 × 800 pixelov
Zväčšenie	4x, 10x, 20x, 40×
Formát	JPG farebné snímky
Typ tkaniva	Tkanivo hrubého čreva

Tabuľka 2: Špecifikácia CRC-HGD-v1 datasetu

2.0.3 Delenie Dát

Dáta boli rozdelené nasledovne s použitím metódy `train_test_split`:

Sada	Počet vzoriek	Podiel	Zdravé/Nádorové
Trénovacia (Train)	8 000	80%	4 000 / 4 000
Validačná (Val)	1 000	10%	500 / 500
Testovacia (Test)	1 000	10%	500 / 500
Celkem	**10 000**	**100%**	**5 000 / 5 000**

Tabuľka 3: Delenie dát na train/val/test sady

3 Použité technológie

3.1 Preprocessing

Zo stredovej časti snímky som vybral 50% tak, aby som používal len reálne tkanivo a vynechal výrez snímky vytvorený mikroskopom. Zároveň som v preprocessingu použil aj metodu rgb2gray na prevedenie snímok z rgb formátu na čiernobiely. To zaistí konštatnejšie farby aj pri snímkach z rôznych datasetov.

3.2 GLCM - Gray Level Co-occurrence Matrix

GLCM metóda je klasickým prístupom pre extrakciu štatistických textúrnych deskriptorov. Zachytáva priestorový vzťah medzi pixelmi s podobnými intenzitami.

Princíp: Matica GLCM počíta výskyt párov pixelov s určitými intenzitami na určitej vzdialenosti a smere.

Parametre použité:

- Úrovne sivej: 32 (quantization levels)
- Vzdialenosti: [1, 2, 5] pixelov
- Uhly: $0, 45, 90, 135$ (4 smery)

Počet prvkov: Pre 4 smery a 3 vzdialenosti s 5 vlastnosťami: $4 \times 3 \times 5 = 60$ prvkov.

3.3 LBP - Local Binary Pattern

LBP je metóda na detekovanie lokálnych textúr. Kóduje binárny vzor okolo každého pixelu v porovnaní s ním samým.

Princíp: Pre každý pixel sa porovnajú hodnoty jeho susedných pixelov so stredom. Vznikne 8-bitový binárny kód, ktorý sa interpretuje ako desatinné číslo.

Výstup: Histogram jednotlivých LBP kódov. Dokopy dostaneme 10 dimenzií.

3.4 GLRLM - Gray Level Run Length Matrix

GLRLM je aproximácia metódou detekcie čiar. Reprezentuje dĺžku sekvencií pixelov s rovnakou intenzitou v špecifických smeroch.

Implementácia: Namiesto úplného GLRLM algoritmu som použili proxy na základe Sobelovho filtra (detekcia hrán). Histogram gradientov je rozdelený do 32 odtieňov sivej.

Výstup: **32 dimenzií**.

3.5 Kombinovaný Vektor Deskriptorov

Všetky tri metódy sa spojili do jedného feature vektora:

$$\mathbf{x} = [\text{GLCM}_{60D} \oplus \text{LBP}_{10D} \oplus \text{GLRLM}_{32D}] \tag{1}$$

Celkový počet prvkov: $60 + 10 + 32 = 102$ dimenzií

Normalizácia: Každý vektor bol normalizovaný pomocou L2 normy na jednotkový vektor:

$$\mathbf{x}_{\text{norm}} = \frac{\mathbf{x}}{\|\mathbf{x}\|_2} = \frac{\mathbf{x}}{\sqrt{\sum_{i=1}^{102} x_i^2}} \tag{2}$$

3.6 Modely Strojového Učenia

Natrénoval som 4 rôze modely s cieľom porovnať ich výkon:

3.6.1 Support Vector Machine (SVM)

Typ: SVC s RBF kernelom

Parametre:

```
C = 10
kernel = 'rbf'
gamma = 'scale'
probability = True
```

3.6.2 Random Forest (RF)

Typ: Súbor náhodných rozhodovacích stromov

Parametre:

```
n_estimators = 200
max_depth = 10
random_state = 42
```

3.6.3 XGBoost (XGB)

Typ: Extreme Gradient Boosting

Parametre:

```
n_estimators = 200
max_depth = 6
random_state = 42
eval_metric = 'logloss'
```

3.6.4 Gradient Boosting (GB)

Parametre:

```
n_estimators = 100
max_depth = 3
learning_rate = 0.1
random_state = 42
```

4 Výsledky

4.1 Výkon Modelov na LC25000

4.1.1 Porovnanie AUC Skóre

Všetky modely dosiahli výnimočný výkon na testovacej sade:

Model	AUC	F1-Macro	Presnosť	Ranking
SVM	0.9836	0.9404	0.9405	3.
RF	0.9830	0.9392	0.9395	4.
XGB	0.9975	0.9765	0.9765	1.
GB	0.9957	0.9690	0.9690	2.

Tabuľka 4: Porovnanie výkonu všetkých modelov na testovacej sade

Záver: XGB dosahuje najvyšší AUC (0.9975). Všetky modely operujú na veľmi vysokej úrovni. Rozdiel medzi najlepším a najhorším je iba 0.0145 (1.45 percentného bodu).

4.2 Detailná Analýza XGB Modelu

XGB model sa ukázal ako najlepší.

	Predikov. Zdravé	Predikoved. Tumor
Skutočne Zdravé	491 (TN)	9 (FP)
Skutočne Tumor	16 (FN)	484 (TP)

Tabuľka 5: Matica zámeny XGB modelu na testovacej sade

4.2.1 Confusion Matrix

Legenda:

- **TN** (True Negative) = 491: Zdravé vzorky správne označené ako zdravé
- **FP** (False Positive) = 9: Zdravé vzorky chybne označené ako tumor
- **FN** (False Negative) = 16: Tumorózne vzorky chybne označené ako zdravé
- **TP** (True Positive) = 484: Tumorózne vzorky správne označené ako tumor

4.2.2 Senzitivita

Senzitivita meraní, aký podiel skutočných tumorov bol správne detekovaný:

$$\text{Senzitivita} = \frac{TP}{TP + FN} = \frac{491}{491 + 9} = \frac{491}{500} = 0.982 = 98.2\% \tag{3}$$

Interpretácia: Z 500 pacientov so skutočným nádorom by bol 491 správne diagnostikovaný a 9 by bolo prehliadnutých.

Podiel zdravých vzoriek, ktoré boli správne identifikované:

$$\text{Špecifičnosť} = \frac{TN}{TN + FP} = \frac{484}{484 + 16} = \frac{484}{500} = 0.968 = 96.8\% \tag{4}$$

Interpretácia: Z 500 pacientov, ktorí sú skutočne zdraví, by bol 484 správne označený ako zdravý a 16 by nesprávne chorých.

4.2.3 Presnosť - Precision

Precision meraní, ako často je model správny, keď predikuje nádor:

$$\text{Presnosť} = \frac{TP}{TP + FP} = \frac{484}{484 + 9} = \frac{484}{493} \approx 0.9817 = 98.2\% \tag{5}$$

Interpretácia: Keď model predikuje tumor, je to správne v 98.2% prípadoch. Toto je dôležité pre zabránenie zbytočným terapiám.

4.2.4 Presnosť - Accuracy

Celková presnosť meraní podiel správnych prediktí z celkového počtu:

$$\text{Accuracy} = \frac{TP + TN}{\text{Celkem}} = \frac{484 + 491}{1000} = \frac{975}{2000} = 0.975 = 97.5\% \quad (6)$$

4.3 Reprodukovateľnosť Výsledkov

Všetky modely dosiahli veľmi podobné výsledky pri viacerých spusteniach. To je kvôli:

5 Diskusia

Úspešnosť modelov pri testovaní na rovnakom datasete na akom boli učené je veľmi dobrá. Na rozdiel od toho ked použijem druhý dataset na Transfer testing, všetky snímky sú označené ako zdravé a nepodarilo sa mi to napraviť. Dané dva datasety sú pravdepodobne príliš odlišné konkrétnym tkanivom, priblížením, sýtosťou farieb ale aj rozlíšením. To zapričiňuje nefunkčnosť mojich natrénovaných modelov na datasete LC25000. Keďže mám v datasete snímky v rôznych priblíženiach, skúšal som či pomôže vyskúšať transfer testing s takouto zmenou ale rovnako neúspešne.

5.1 Porovnanie Rôznych Modelov

5.1.1 Výsledky

Poradie modelov podľa AUC (v priemere):

1. **XGB**: 0.99 (Gradient Boosting - alternatíva)
2. **GB**: 0.99 (Klasické Boosting)
3. **SVM**: 0.98 (Kernel Methods)
4. **RF**: 0.97 (Ensemble Shallow)

6 Záver

V tejto práci som vyvinul a evauloval systém klasifikácie medicínskych snímkov tkaniva hrubého čreva. Môj prístup kombinuje:

1. **Klasické textúrne charakteristiky**: GLCM, LBP, GLRLM vytvárajúce 102-dimenzionálny feature vektor
2. **4 algoritmy strojového učenia**: SVM, Random Forest, XGBoost, Gradient Boosting
3. **Kompletnú evaluáciu**: Confusion matrix, senzitivita, špecifickosť, presnosť, AUC

6.1 Implikácie

Textúrne deskriptory v kombinácií so strojovým učením sú účinnou alternatívou k ľudkej evaluácii. Moja implementácia by mohola slúžiť ako:

- Podpornú nástroj pre patológov pri diagnostike
- Automatizovaný screening systém pred endoskopickou procedúrou
- Výukový nástroj pre standardizáciu v medicínskom vzdelávaní