**Т Е Х Н И Ч Е С К И У Н И В Е Р С И Т Е Т – С О Ф И Я**

**КАТЕДРА “ИНФОРМАТИКА”**

**СПЕЦИАЛНОСТ “АНАЛИЗ НА ГОЛЕМИ МАСИВИ И ПОТОЦИ ОТ ДАННИ”**

**Степен Магистър**

**“Анализ на големи данни за прецизната медицина”**

**ТЕМА**

**“РАК НА БЕЛИЯ ДРОБ”**

**Разработил:**

1. **Анна- Мария Атанасова 961323002**

**2024 г.**

[**1. Обща информация за заболяването с фокус върху социалната му значимост в световен и национален мащаб. 4**](#_lepyxjfpsskx)

[Видове рак на белия дроб 4](#_h39y8rw433tm)

[Рискови фактори и причини 4](#_g50f1r9pzi1t)

[Симптоми и диагноза 4](#_tbgr2lvtw7rr)

[Опции за лечение 4](#_18vt3yz59tn8)

[Продължителност на живота при различните случаи 5](#_oxajiq5b9udi)

[Глобално въздействие на рака на белия дроб 5](#_mt38matmpx41)

[Рак на белия дроб в България 5](#_awmgr79wjdej)

[Социалното значение на рака на белия дроб в България 6](#_3o44mjwrensu)

[Обществени здравни инициативи и предизвикателства 6](#_cteoeoymv6qr)

[**2. Описание на набора данни за диагностичния анализ. 6**](#_1w1i16tf5vwc)

[2А. Описание на характеристиките 7](#_lyaqlnxu7l4c)

[2Б. Статистически анализ на характеристиките – визуализация, корелационна матрица (heat map), и др. 7](#_xof1f7yq0n3f)

[Визуализация на част от изображенията с техния label 7](#_24iae65l8jc5)

[Разпределение на изображенията по Labels 8](#_ah5bah35lbf)

[Heatmap 9](#_2v8krhx9d550)

[Големина на изображенията 9](#_e8nhntc7qxnc)

[В байтове 10](#_xl4ympfy35y3)

[В мегабайти 10](#_cli5c6m36s8r)

[Резолюция на изображенията 11](#_rbb792u9p2ck)

[Разпределение на цветовете в изображенията 11](#_3mva69hr6wui)

[2В. Предварителна обработка на набора данни (preprocessing) – зашуменост, липсващи данни, редукция на размерност, и др. 12](#_jrsr8qod8q8c)

[Липсващи данни 12](#_5k1osfj2zs6n)

[Редукция на размера на изображенията на 224х224 пиксела 12](#_a6k5dr5a6psv)

[**3. Технологична работна рамка – програмна среда, език за програмиране, библиотеки, и др. 12**](#_xajgrmi9v84)

[Програмна среда 12](#_639trqvwx9g3)

[Език за програмиране 12](#_p0zzmaw1nik)

[Библиотеки 12](#_fgv2ra4pl654)

[Библиотеки 12](#_urbc0ee5vr5b)

[Пакети на трети страни 12](#_e2da66b33wx7)

[**4. Описание на двата алгоритъма за класификация. 13**](#_cvzb3065664d)

[Метод на Опорните Вектори (SVM): 13](#_8m0hakqmshto)

[Подготовка на данните 13](#_dhoddtw0qvxp)

[Извличане на характеристики 13](#_u08g7nf94tp)

[Трениране на модела 13](#_v6mvq4eakaz)

[Оценка на модела 13](#_4swhrljfce8c)

[Конволюционна Невронна Мрежа (CNN): 13](#_m5f48rbhax19)

[Подготовка на данните 13](#_rp56got0sd9p)

[Създаване на модела 13](#_ibe7kfkuoa60)

[Компилация и трениране на модела 13](#_yn6g1lfujqlv)

[Оценка на модела 13](#_fmdb3jmk76)

[**5. Синтез на моделите за машинно обучение за диагностичен анализ – конфигуриране на REPO в GitHub със сорс кодовете на двата модела и линк към него. 14**](#_emulitf6y7mv)

[**6. Експериментални резултати 14**](#_9iy8fgmbqcg8)

[6А. Скрийншотове от обучението и валидирането на двата модела. 14](#_jdzsarpltajj)

[SVM model 14](#_4kb9qgt2epjg)

[CNN model 15](#_r91vmdefupd3)

[6Б. Сравнителни диаграми за двата алгоритъма на оценките на параметрите PAS (Precision, Accuracy, Sensitivity) 15](#_95hue67g6qbm)

[6В. Сравнителни диаграми за двата алгоритъма на оценките на параметрите FP, FN, TP, TN 17](#_hik4862a3o0n)

[**7. Информационни ресурси. 17**](#_i5a2yo9uhfix)

# 

# 1. Обща информация за заболяването с фокус върху социалната му значимост в световен и национален мащаб.

Ракът на белия дроб остава едно от най-разпространените и смъртоносни форми на рак в световен мащаб. Характеризира се с неконтролиран растеж на клетки в тъканите на белия дроб и често остава недиагностициран, докато не достигне напреднал стадий, което прави лечението по-трудно и намалява шансовете за оцеляване.

## Видове рак на белия дроб

Ракът на белия дроб се категоризира в две основни групи: недребноклетъчен рак на белия дроб (NSCLC) и дребноклетъчен рак на белия дроб (SCLC). NSCLC представлява приблизително 85% от всички случаи на рак на белия дроб и включва подтипове като аденокарцином, плоскоклетъчен карцином и едроклетъчен карцином. SCLC, съставляващ около 15% от случаите, е по-агресивен и се разпространява по-бързо от NSCLC.

## Рискови фактори и причини

Основната причина за рак на белия дроб е тютюнопушенето, отговорно за около 85% от всички случаи. Това включва както активно пушене, така и излагане на вторичен дим. Други значими рискови фактори включват излагане на радон газ, азбест, замърсяване на въздуха и професионални рискове като определени химикали и прах. Генетичната предразположеност и семейната история на рак на белия дроб също могат да увеличат риска.

## Симптоми и диагноза

Ранните стадии на рак на белия дроб често не показват симптоми, което допринася за късното му диагностициране. Когато се появят симптоми, те могат да включват постоянна кашлица, кашляне на кръв, болка в гърдите, задух, необяснима загуба на тегло и повтарящи се респираторни инфекции. Диагностичните методи включват образни изследвания като рентгенови лъчи и компютърна томография, цитология на храчките и биопсия на тъканите.

## Опции за лечение

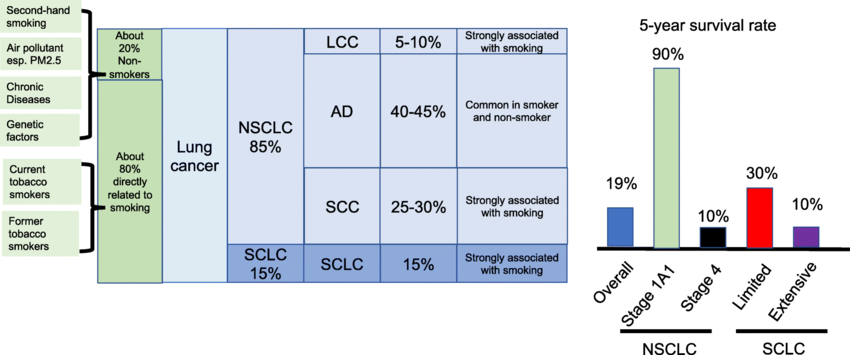
Лечението на рака на белия дроб зависи от типа и стадия на заболяването. Опциите включват хирургия, лъчетерапия, химиотерапия, таргетна терапия и имунотерапия. Ранните стадии на NSCLC могат да бъдат лекувани ефективно с хирургия, докато напредналите стадии често изискват комбинация от лечения. SCLC, поради своята агресивна природа, обикновено реагира по-добре на химиотерапия и лъчетерапия.

## Продължителност на живота при различните случаи

Общата 5-годишна преживяемост при рак на белия дроб е около 19%. Ниските нива на преживяемост при рак на белия дроб се дължат на пациенти, диагностицирани с метастатично заболяване.

Сред пациентите с NSCLC 5-годишната относителна преживяемост достига 90% за стадий 1A1, но пада под 10% за стадий 4.

Сред пациентите с SCLC 5-годишната относителна преживяемост е около 30% за ограничено заболяване и под 10% за екстензивно заболяване.



Фиг 1. Дребноклетъчен белодробен карцином (SCLC), недребноклетъчен белодробен карцином (NSCLC), плоскоклетъчният карцином (SCC), аденокарциномът (AD) и едроклетъчният карцином (LCC)

## Глобално въздействие на рака на белия дроб

На глобално ниво, ракът на белия дроб е водещата причина за смърт от рак, с приблизително 1.8 милиона смъртни случая през 2020 г., което представлява 18% от всички смъртни случаи от рак. Високата смъртност се дължи на тенденцията заболяването да бъде диагностицирано в напреднал стадий и ограничената наличност на програми за ранно откриване в много региони. Освен това, икономическата тежест на рака на белия дроб е значителна, включваща директни медицински разходи, загубена продуктивност и дългосрочни разходи за инвалидност.

## Рак на белия дроб в България

В България, ракът на белия дроб представлява значим обществен здравен проблем. Според Световната здравна организация (СЗО), България има една от най-високите смъртности от рак на белия дроб в Европа. Разпространението на тютюнопушенето сред българското население е основен допринасящ фактор. Доклад от Националния статистически институт (НСИ) на България разкрива, че над 30% от българските възрастни са пушачи, един от най-високите проценти в Европейския съюз.

## Социалното значение на рака на белия дроб в България

Социалното значение на рака на белия дроб в България е дълбоко. Високата разпространеност и смъртност засягат не само здравната система, но и по-широкото общество. Семействата и общностите често са засегнати от загубата на продуктивни членове, а емоционалният и финансовият натиск върху семействата е значителен. Освен това, здравната система се сблъсква с предизвикателства при предоставянето на адекватни ресурси за превенция, ранно откриване и лечение на рак на белия дроб.

## Обществени здравни инициативи и предизвикателства

Усилията за борба с рака на белия дроб в България включват обществени здравни кампании, насочени към намаляване на тютюнопушенето, подобряване на качеството на въздуха и насърчаване на ранното откриване чрез редовни скрининги. Българското правителство е въвело различни закони и регулации против тютюнопушенето, включително забрани за реклама, ограничения за пушене на обществени места и повишени данъци върху тютюневите изделия. Въпреки това, прилагането на тези закони остава предизвикателство, а процентите на пушене, особено сред по-младите популации, остават високи.

Здравната инфраструктура в България също се сблъсква със значителни предизвикателства. Има неравенства в достъпа до здравни услуги между градските и селските райони, като населението в селските райони често има ограничен достъп до специализирана медицинска помощ. Освен това, има нужда от увеличаване на инвестициите в здравни заведения, обучение на медицински професионалисти и обществени информационни кампании за образование на населението относно рисковете от рак на белия дроб и важността на ранното откриване.

# 2. Описание на набора данни за диагностичния анализ.

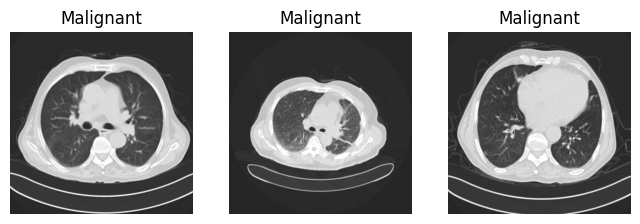
Наборът от данни за рак на белия дроб е събран за период от 3 месеца през есента на 20219г от Иракска онкологична болница/Национален център за ракови заболявания. Съдържа компютърна томография на пациенти, диагностицирани с белодробно заболяване рак в различни стадии, както и здрави лица. Наборът от данни съдържа общо 1097 изображения, представляващи срезове от компютърна томографии на 110 случая. Тези случаи са групирани в три класа: нормални, доброкачествени и злокачествени. от тях 40 случая са диагностицирани като злокачествени; 15 случая са диагностицирани като доброкачествени и 55 случая са класифицирани като нормални случаи.

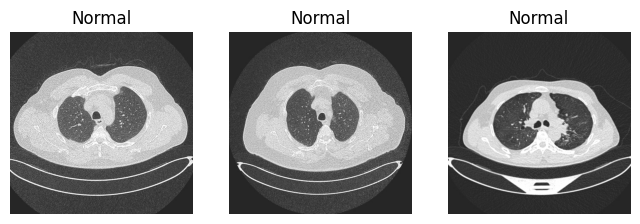
## 2А. Описание на характеристиките

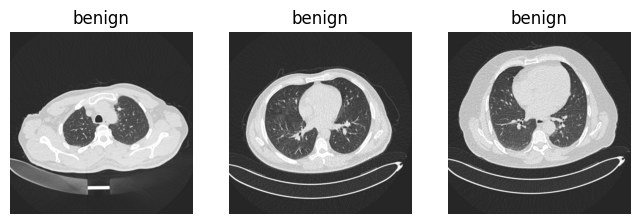
* Filepaths - адреси на изображения с формат .jpg, представляващи срезове от компютърна томографии на 110 случая.
* Labels - описващ вид/наличие на рак. Рапределят се на “Malignant” - 561 записи, “Normal” - 416 записи и “benign” - 120 записи

## 2Б. Статистически анализ на характеристиките – визуализация, корелационна матрица (heat map), и др.

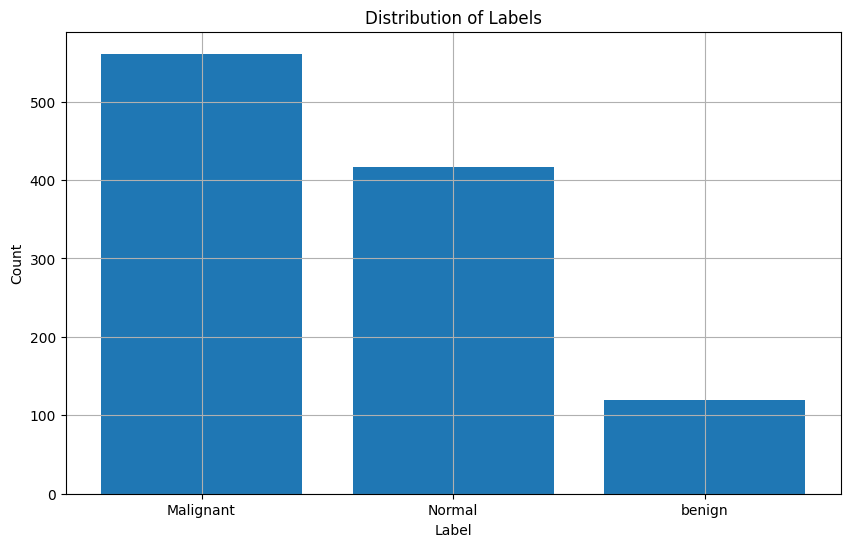
### Визуализация на част от изображенията с техния label







### Разпределение на изображенията по Labels

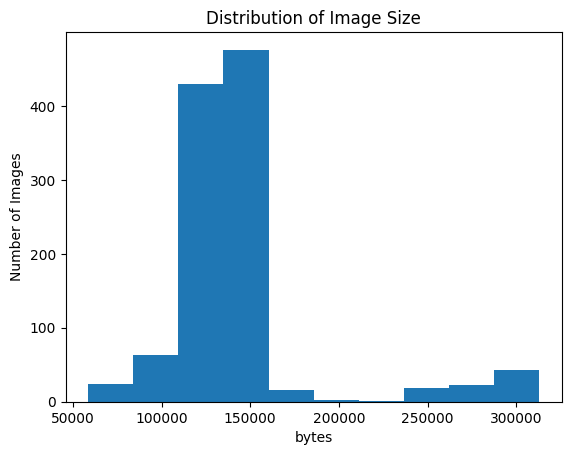


### Heatmap

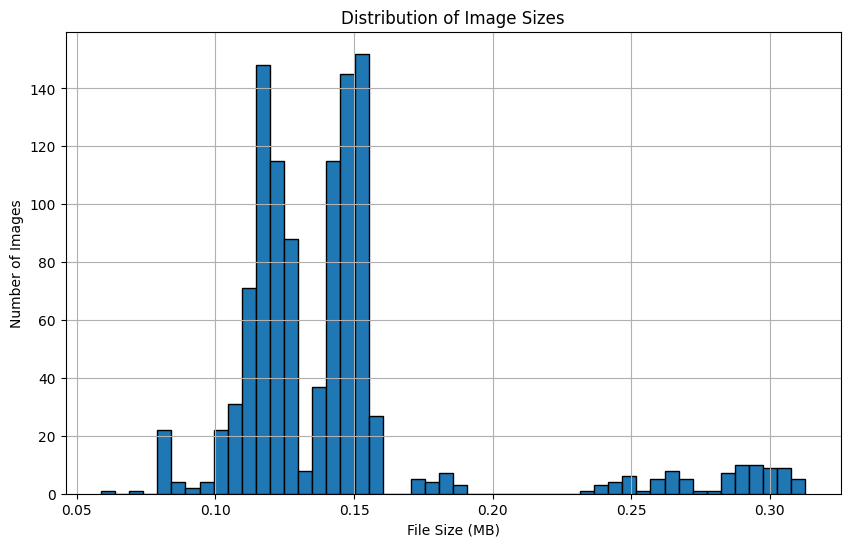
### 

### Големина на изображенията

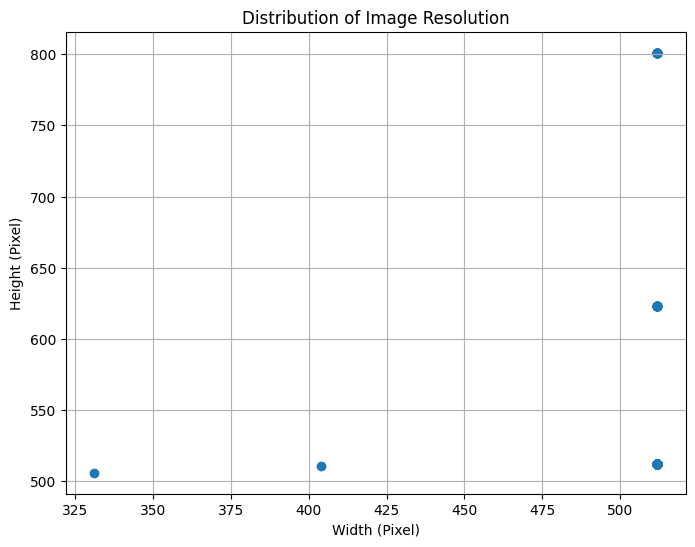
#### В байтове



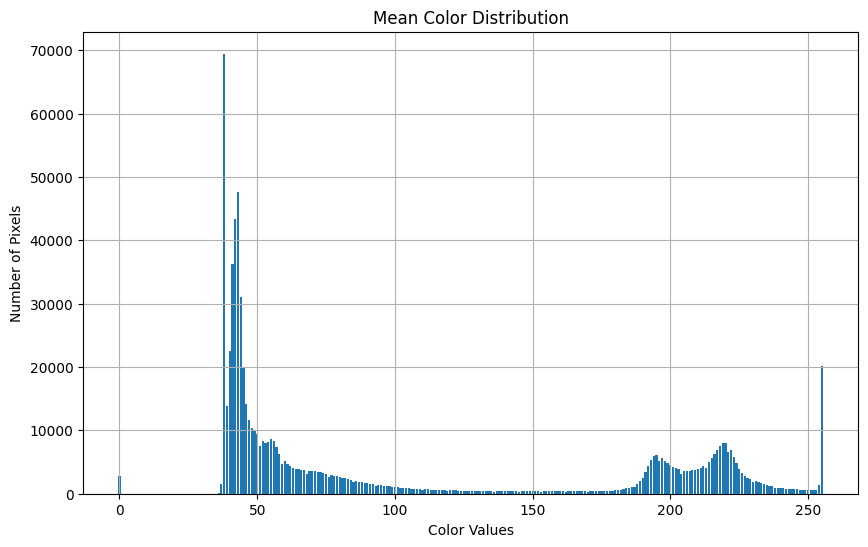
#### В мегабайти



## Резолюция на изображенията

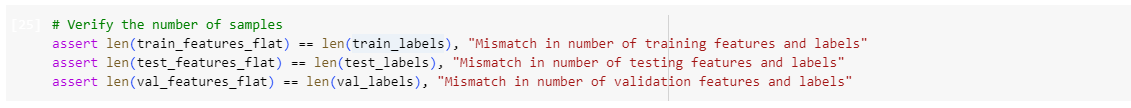


## Разпределение на цветовете в изображенията



## 2В. Предварителна обработка на набора данни (preprocessing) – зашуменост, липсващи данни, редукция на размерност, и др.

### Липсващи данни



### Редукция на размера на изображенията на 224х224 пиксела

# 3. Технологична работна рамка – програмна среда, език за програмиране, библиотеки, и др.

## Програмна среда

Google Colab

## Език за програмиране

Python

## Библиотеки

### Библиотеки

* Google.colab.drive
* os

### Пакети на трети страни

* Cv2
* Pandas
* Seaborn
* Matplotlib.pyplot
* Numpy
* Tensorflow
* PIL.Image
* sklearn

# 4. Описание на двата алгоритъма за класификация.

## Метод на Опорните Вектори (SVM):

### Подготовка на данните

Данните за обучение и тестваме се зареждат и се предоставят на ImageDataGenerator за предварителна обработка и аугментация.

### Извличане на характеристики

Използва се предварително обучена невронна мрежа (MobileNetV2) за извличане на характеристики от изображенията.

### Трениране на модела

Характеристиките от изображенията се използват като входни данни за трениране на SVM модел, който се обучава да класифицира изображенията в съответните класове.

### Оценка на модела

Моделът се оценява чрез предсказване на класовете за тестовия набор и сравняване на предсказанията с истинските класове.

## Конволюционна Невронна Мрежа (CNN):

### Подготовка на данните

Също като за SVM, данните се зареждат и предоставят на ImageDataGenerator за предварителна обработка.

### Създаване на модела

Създаваме Sequential модел на Keras, който се състои от няколко слоя за конволюция и пулинг, последвани от пълно свързани слоеве за класификация.

### Компилация и трениране на модела

Моделът се компилира с оптимизатор и функция за загуба, след което се тренира върху обучаващия набор с помощта на функцията fit.

### Оценка на модела

Моделът се оценява чрез използване на тестовия набор, като се изчислява точността на класификацията.

# 5. Синтез на моделите за машинно обучение за диагностичен анализ – конфигуриране на REPO в GitHub със сорс кодовете на двата модела и линк към него.

<https://github.com/an00000000na/Lung-Cancer.git>

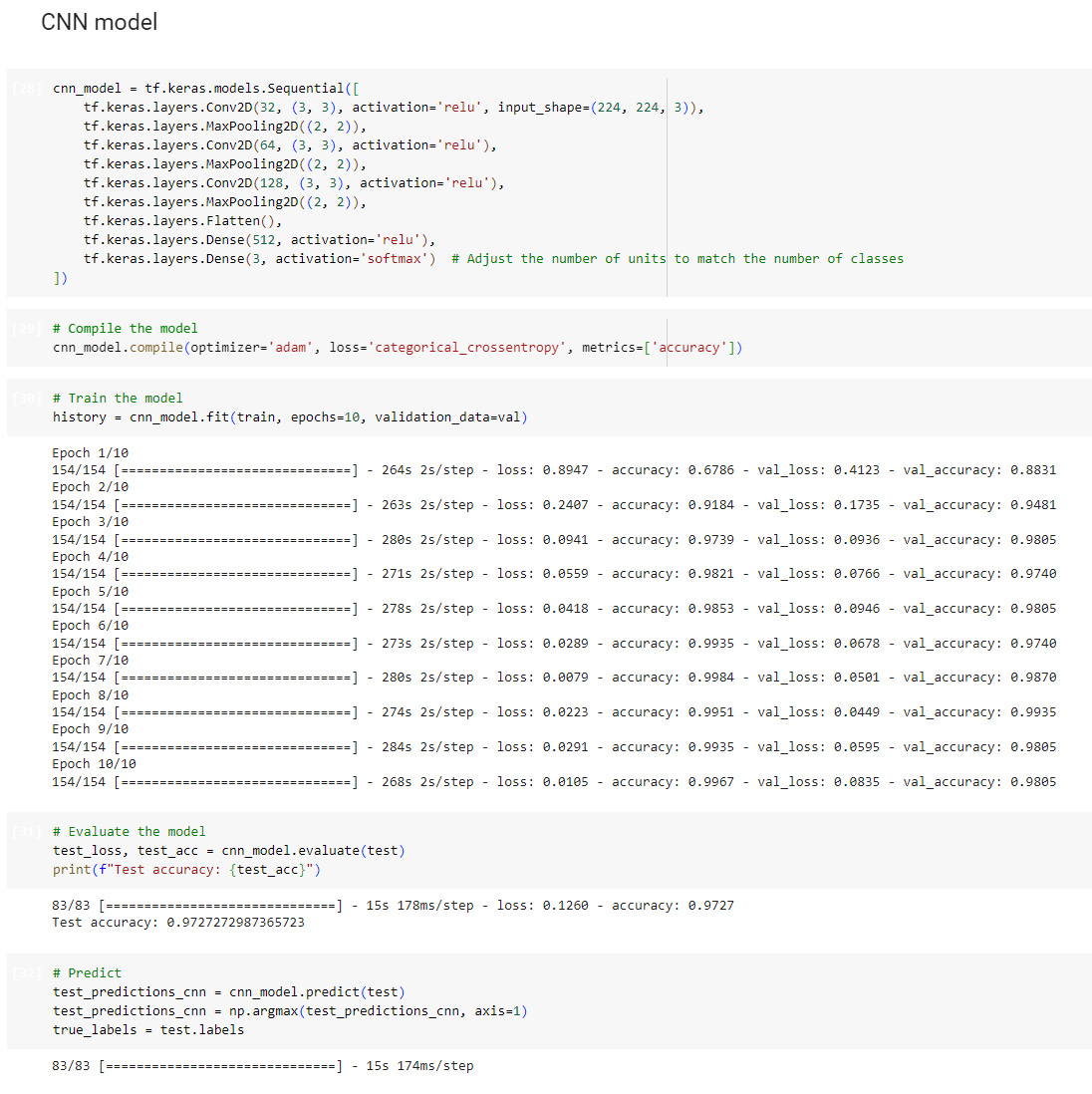
# 6. Експериментални резултати

## 6А. Скрийншотове от обучението и валидирането на двата модела.

### SVM model



### CNN model



## 6Б. Сравнителни диаграми за двата алгоритъма на оценките на параметрите PAS (Precision, Accuracy, Sensitivity)

SVM Model PAS Metrics:

Precision (Macro-Averaged): 0.9735

Recall (Macro-Averaged): 0.9167

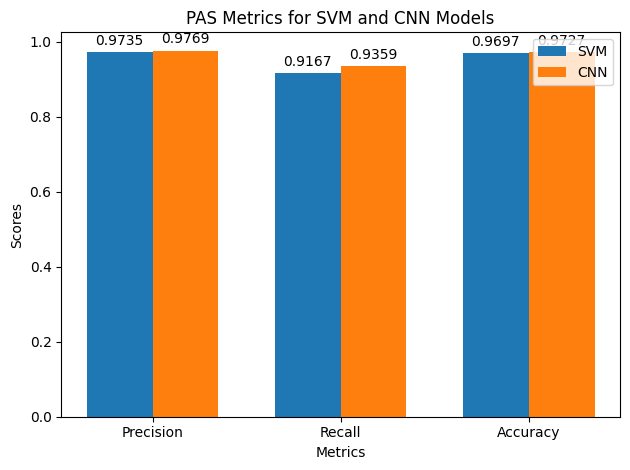
Accuracy: 0.9697

CNN Model PAS Metrics:

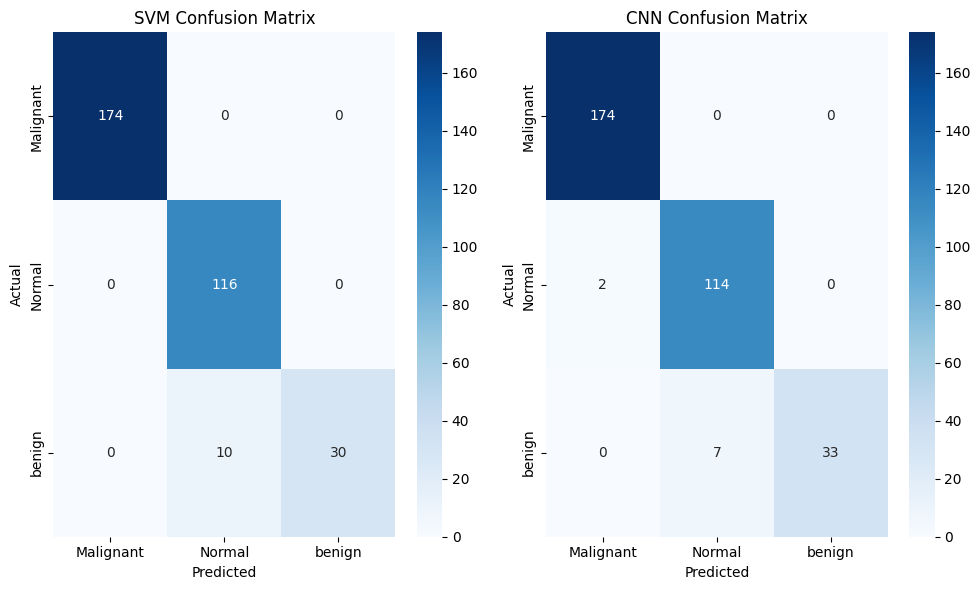
Precision (Macro-Averaged): 0.9769

Recall (Macro-Averaged): 0.9359

Accuracy: 0.9727



## 6В. Сравнителни диаграми за двата алгоритъма на оценките на параметрите FP, FN, TP, TN



# 7. Информационни ресурси.

1. Набор данни: <https://www.kaggle.com/datasets/adityamahimkar/iqothnccd-lung-cancer-dataset/data>
2. American Cancer Society. (2020). Types of Lung Cancer. <https://www.cancer.org/cancer/lung-cancer/about/what-is.html>
3. Mayo Clinic. (2021). Lung Cancer Symptoms and Causes. <https://www.mayoclinic.org/diseases-conditions/lung-cancer/symptoms-causes/syc-20374620>
4. Very Well Health - Stage 4 Lung Cancer Life Expectancy <https://www.verywellhealth.com/what-is-stage-4-lung-cancer-life-expectancy-2249420>
5. Progress on pivotal role and application of exosome in lung cancer carcinogenesis,diagnosis, therapy and prognosis - Ming-Yue Li, Li-Zhong Liu and Ming Dong <https://www.researchgate.net/publication/348816466_Progress_on_pivotal_role_and_application_of_exosome_in_lung_cancer_carcinogenesis_diagnosis_therapy_and_prognosis>