

**Schilderijen classifcatie**

student **Tim Mortier**

studiegebied **handelswetenschappen en bedrijfskunde**

bachelor in de **Toegepaste Informatica**

campus **Kortrijk**

academiejaar **2024-2025**

docent **A. Louwyck**

## Omgeving

Ik heb een GitHub-repository opgezet waarin ik mijn notebooks en data opsla. Mijn laptop beschikt over een GPU met 8 GB RAM, waarmee ik eenvoudige neurale netwerken kan trainen. Deze hardware is echter niet voldoende voor het trainen van grotere neurale netwerken, waarvoor een GPU met 16 GB RAM nodig is. Hierdoor maakte ik gebruik van Kaggle, waar ik 30 uur per week toegang heb tot een GPU met 16 GB aan RAM. Via Kaggle was het mogelijk om mijn GitHub-repository te koppelen, zodat ik notebooks kon pullen en pushen.

## Data cleaning

Ik voegde de schilderijen van de 3 andere schilders toe aan mijn zelf gescrapte schilderijen van rembrandt. Ik onderzocht eerst als er corrupte afbeeldingen in de dataset zaten, maar dat was niet het geval.  
Vervolgens keek ik als de formaten van de afbeeldingen wel allemaal RGB en JPG waren. Er waren een aantal JPG afbeeldingen die niet RGB waren en ook 2 PNG afbeeldingen. Ik zette deze om naar RGB en JPG.

Hierna nummerde ik de schilderijen per schilder.

Ik bekeek ook de aantal schilderijen per schilder en zag dat de dataset niet belanced was. Zo zijn er vijf keer meer schilderijen van Picasso dan Mondriaan in de dataset. Hierdoor besloot ik een imbalanced, undersampled en oversampled dataset te creëren om later te onderzoeken welke samplingstrategie het best gebruikt wordt.

## EDA

Afbeelding met tekst, schermopname, diagram, Perceel

Automatisch gegenereerde beschrijving

Ik onderzocht de hoogte en breedte van de afbeeldingen. Ik zag dat het grootste deel van de afbeeldingen een hoogte en breedte had tussen 100 en 2000 pixels. Er zijn ook enkele outliers die tot boven de 7000 pixels gaan.

Afbeelding met tekst, schermopname, diagram, Perceel

Automatisch gegenereerde beschrijving

In de histogrammen van de hoogtes en breedtes valt ook aan de frequentie te zien dat de dataset niet gebalanceerd is. Rembrandt heeft afbeeldingen van (min) tot (max). Mondriaan van (min) tot (max). Rubens van (min) tot (max). Picasso van (min) tot (max). Ik sla de algemene statistieken op om later te bekijken welke hoogte en breedte het best gebruikt worden in het model.

SCHILDERIJEN

## Data sampling and reshaping

In deze notebook pas ik verschillende samplingstrategieën toe en t bepaal ik wat de optimale hoogtes en breedtes zijn. Voor de evaluatie van de modellen besloot ik de dataset te splitsen volgens hold-out validation. Ik hanteerde een verdeling van 60% trainingsdata, 20% validatiedata en 20% testdata. K-fold crossvalidatie koos ik niet, omdat de ongebalanceerde dataset en de oversampled dataset honderden afbeeldingen per schilder bevat. Bij de undersampled dataset zijn er echter slechts 52 afbeeldingen in de validatie- en testset. Daarom besloot ik eenvoudige data-augmentatie toe te passen en telkens het verschil te analyseren tussen datasets met en zonder data-augmentatie. In een later stadium onderzoek ik welke data-augmentatiestrategie het meest effectief is.

Ik start met de classificatie van werken van twee schilders. Daarna onderzoek ik of dezelfde technieken toepasbaar zijn op meerdere schilders. Ik testte twee paren schilders: Mondriaan en Picasso, en Rubens en Rembrandt.

Voor het trainen gebruikte ik een eenvoudig neuraal netwerk gebaseerd op listing XX uit het boek. Dit model behaalde vaak slechts baseline accuracies op de testset, terwijl de trainings- en validatie-accuracies tussen 70% en 90% schommelden. Om overfitting tegen te gaan, voegde ik dropout-layers toe aan het model, maar dit leverde onvoldoende verbetering op. Dit is waarschijnlijk te wijten aan de eenvoud van het model. Om die reden baseerde ik mijn conclusies voornamelijk op de validatie-accuracies.

Bij het kiezen van de sampling strategie koos ik mediaanhoogte en -breedte als initiële shape om een gemiddeld resultaat te verkrijgen. De keuze voor de mediaan boven het gemiddelde werd gemaakt omdat de mediaanwaarde daadwerkelijk voorkomt in de dataset.

### Undersampled dataset

#### Zonder data augmentatie

Afbeelding met tekst, diagram, lijn, Perceel

Automatisch gegenereerde beschrijvingHet model heeft hier een accuracy van 46% wat lager is dan de baseline van 50%. Het model heeft een recall van 21.57%, het voorspelt dus dit percentage van klasse 1 daadwerkelijk als klasse 1. De precision is 42.31%, dit percentage aan voorspelde klasse 1 schilderijen is daadwerkelijk ook klasse 1. Het harmonisch gemiddelde van de laatse twee metrieken is 28.57%. ROC AUC

#### Met data augmentatie

Afbeelding met tekst, diagram, lijn, schermopname

Automatisch gegenereerde beschrijving

Het model voorspelt met data augmentatie duidelijker al wat beter. De accuracy is nu 55% wat hoger is dan de baseline van 50%. Het model heeft een recall van 56.68%, het voorspelt dus dit percentage van klasse 1 daadwerkelijk als klasse 1. De precision is 55.76%, dit percentage aan voorspelde klasse 1 schilderijen is daadwerkelijk ook klasse 1. Het harmonisch gemiddelde van de laatse twee metrieken is 56.31%. ROC AUC.

### Oversampled dataset

#### Zonder data augmentation

Afbeelding met tekst, diagram, lijn, Perceel

Automatisch gegenereerde beschrijving

Het model heeft hier een accuracy van 48% wat lager is dan de baseline van 50%. Het model heeft een recall van 46%, het voorspelt dus dit percentage van klasse 1 daadwerkelijk als klasse 1. De precision is 48%, dit percentage aan voorspelde klasse 1 schilderijen is daadwerkelijk ook klasse 1. Het harmonisch gemiddelde van de laatse twee metrieken is 47%. ROC AUC

#### Met data augmentation

Afbeelding met tekst, diagram, lijn, schermopname

Automatisch gegenereerde beschrijving

Het model heeft hier een accuracy van 50% wat lager is dan de baseline van 50%. Het model heeft een recall van 46%, het voorspelt dus dit percentage van klasse 1 daadwerkelijk als klasse 1. De precision is 48%, dit percentage aan voorspelde klasse 1 schilderijen is daadwerkelijk ook klasse 1. Het harmonisch gemiddelde van de laatse twee metrieken is 47%. ROC AUC