Extraction of Solar Cell Data from PDF Datasheets

Master Thesis Presentation

Muhammad Moeez Malik

Albert-Ludwigs-Universität Freiburg

moeezmalik13@gmail.com

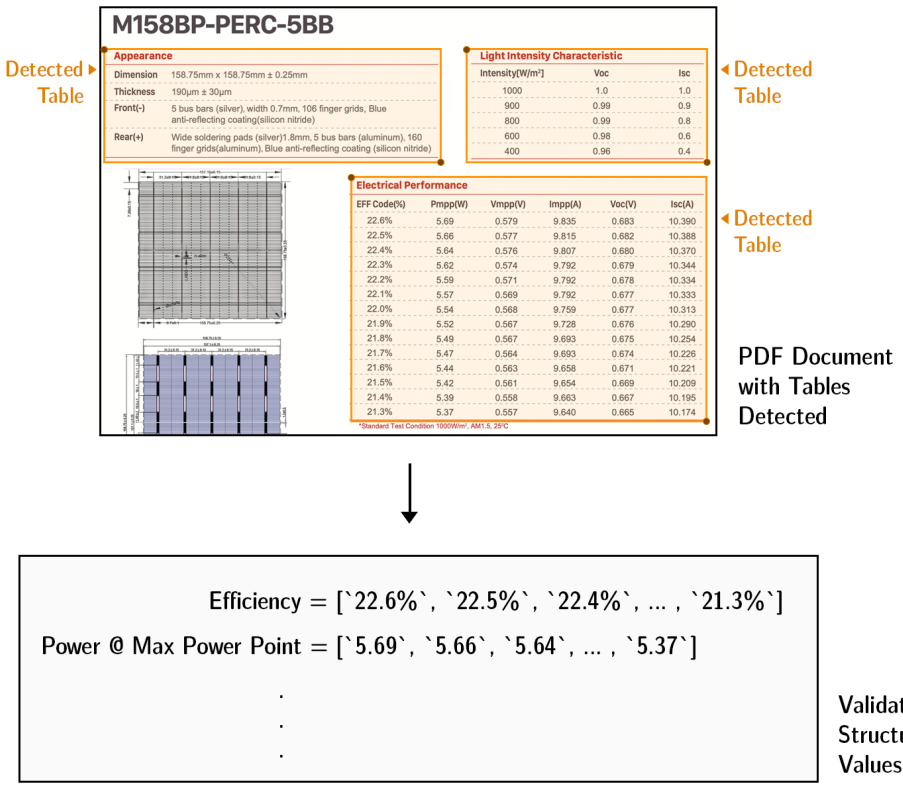
April 14, 2023

Content

Introduction

Introduction

What is the thesis about?



- Automatic extraction of solar cell data from data sheets

Introduction: Motivation

The Need for an Automated Solution

- Automatic generation of indexed and searchable database
- Process a large number of documents quickly
- Prediction of market trends using the data
- Uncovering possible research areas by analysis of data

Introduction: PDF Documents

A Standard for Digital Document Distribution. Why?

- Pros
  - Compact Files
  - Industry standard
  - Same rendering of documents regardless of hardware and software.
  - Ensures information reproducibility
- Cons
  - Document is constructed using only positional information, the row and column information for tables is lost

PDF Document

Introduction: Solar Cell Data Sheets

Distributed as PDF Documents

CONSOLT®

Mono 5BB Solar Cell 158.75 Bifacial

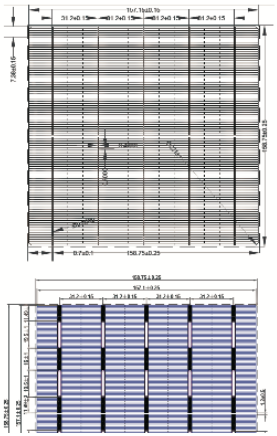
M158BP-PERC-5BB

Appearance

Dimension	158.75mm x 158.75mm ± 0.25mm
Thickness	190μm ± 30μm
Front(-)	5 bus bars (silver), width 0.7mm, 106 finger grids, Blue anti-reflecting coating(silicon nitride)
Rear(+)	Wide soldering pads (silver)1.8mm, 5 bus bars (aluminum), 160 finger grids(aluminum), Blue anti-reflecting coating (silicon nitride)

Light Intensity Characteristic

Intensity[W/m²]	Voc	Isc
1000	1.0	1.0
900	0.99	0.9
800	0.99	0.8
600	0.98	0.6
400	0.96	0.4



Electrical Performance

EFF Code(%)	Pmpp(W)	Vmpp(V)	Impp(A)	Voc(V)	Isc(A)
22.6%	5.69	0.579	9.835	0.683	10.390
22.5%	5.66	0.577	9.815	0.682	10.388
22.4%	5.64	0.576	9.807	0.680	10.370
22.3%	5.62	0.574	9.792	0.679	10.344
22.2%	5.59	0.571	9.792	0.678	10.334
22.1%	5.57	0.569	9.792	0.677	10.333
22.0%	5.54	0.568	9.759	0.677	10.313
21.9%	5.52	0.567	9.728	0.676	10.290
21.8%	5.49	0.567	9.693	0.675	10.254
21.7%	5.47	0.564	9.693	0.674	10.226
21.6%	5.44	0.563	9.658	0.671	10.221
21.5%	5.42	0.561	9.654	0.669	10.209
21.4%	5.39	0.558	9.663	0.667	10.195
21.3%	5.37	0.557	9.640	0.665	10.174

*Standard Test Condition 1000W/m², AM1.5, 25°C

Figure: An example of solar cell data sheet

Introduction: Solar Cell Data Sheets

Distributed as PDF Documents

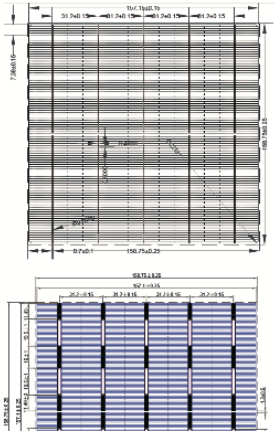
CONSOLT®

Mono 5BB Solar Cell 158.75 Bifacial

M158BP-PERC-5BB

Appearance	
Dimension	158.75mm x 158.75mm ± 0.25mm
Thickness	190μm ± 30μm
Front(-)	5 bus bars (silver), width 0.7mm, 106 finger grids, Blue anti-reflecting coating(silicon nitride)
Rear(+)	Wide soldering pads (silver)1.8mm, 5 bus bars (aluminum), 160 finger grids(aluminum), Blue anti-reflecting coating (silicon nitride)

Light Intensity Characteristic		
Intensity[W/m²]	Voc	Isc
1000	1.0	1.0
900	0.99	0.9
800	0.99	0.8
600	0.98	0.6
400	0.96	0.4



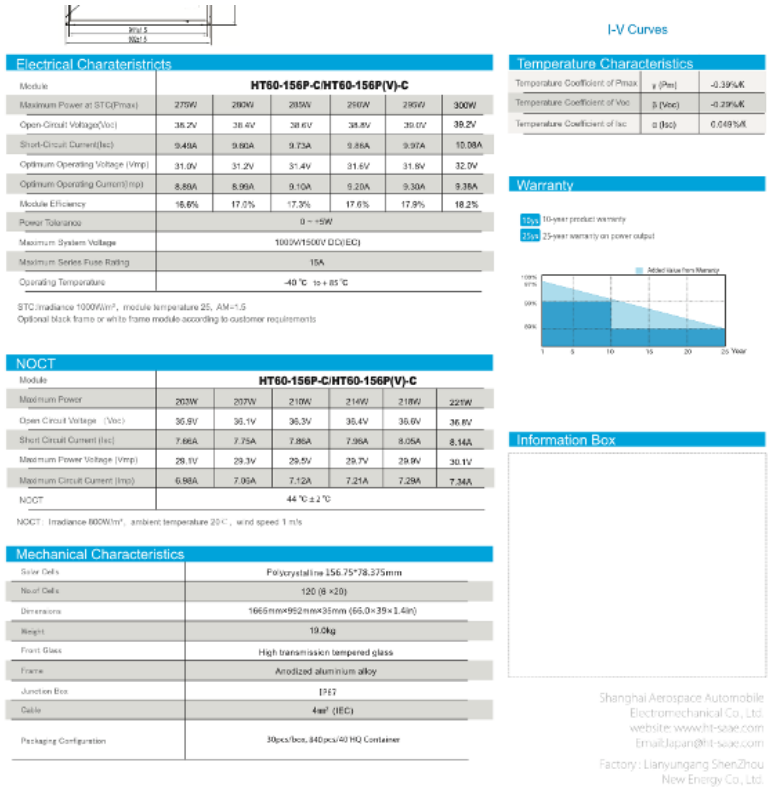
Electrical Performance					
EFF Code(%)	Pmpp(W)	Vmpp(V)	Impp(A)	Voc(V)	Isc(A)
22.6%	5.69	0.579	9.835	0.683	10.390
22.5%	5.66	0.577	9.815	0.682	10.388
22.4%	5.64	0.576	9.807	0.680	10.370
22.3%	5.62	0.574	9.792	0.679	10.344
22.2%	5.59	0.571	9.792	0.678	10.334
22.1%	5.57	0.569	9.792	0.677	10.333
22.0%	5.54	0.568	9.759	0.677	10.313
21.9%	5.52	0.567	9.728	0.676	10.290
21.8%	5.49	0.567	9.693	0.675	10.254
21.7%	5.47	0.564	9.693	0.674	10.226
21.6%	5.44	0.563	9.658	0.671	10.221
21.5%	5.42	0.561	9.654	0.669	10.209
21.4%	5.39	0.558	9.663	0.667	10.195
21.3%	5.37	0.557	9.640	0.665	10.174

*Standard Test Condition 1000W/m², AM1.5, 25°C

Figure: Critical information is in tables

Introduction: Challenges

Variability in Data Sheet Design



I-V Curves

Electrical Characteristics

Module	HT60-156P-C/HT60-156P(V)-C					
Maximum Power at STC(Pmax)	275W	280W	285W	290W	295W	300W
Open-Circuit Voltage(Voc)	38.2V	38.4V	38.6V	38.8V	39.0V	39.2V
Short-Circuit Current(Isc)	9.48A	9.60A	9.73A	9.86A	9.97A	10.08A
Optimum Operating Voltage (Vmp)	31.0V	31.2V	31.4V	31.6V	31.8V	32.0V
Optimum Operating Current(Imp)	8.88A	8.99A	9.10A	9.20A	9.30A	9.38A
Module Efficiency	16.6%	17.0%	17.3%	17.6%	17.9%	18.2%
Power Tolerance	0 ~ +5W					
Maximum System Voltage	1000V/1500V DC(IEC)					
Maximum Series Fuse Rating	15A					
Operating Temperature	-40 °C to +85 °C					

STC:Irradiance 1000W/m², module temperature 25, AM=1.5
Optional black frame or white frame module according to customer requirements

NOCT

Module	HT60-156P-C/HT60-156P(V)-C					
Maximum Power	203W	207W	210W	214W	218W	221W
Open Circuit Voltage (Voc)	35.9V	36.1V	36.3V	36.4V	36.6V	36.8V
Short Circuit Current (Isc)	7.66A	7.75A	7.86A	7.96A	8.05A	8.14A
Maximum Power Voltage (Vmp)	28.1V	28.3V	28.5V	28.7V	28.9V	30.1V
Maximum Circuit Current (Imp)	6.98A	7.05A	7.12A	7.21A	7.29A	7.34A
NOCT	44 °C ± 2 °C					

NOCT: Irradiance 800W/m², ambient temperature 20°C, wind speed 1 m/s

Mechanical Characteristics

Solar Cells	Polycrystalline 156.75*78.375mm
No.of Cells	120 (6 ×20)
Dimensions	1665mm×992mm×35mm (65.0×39×1.4in)
Weight	19.0kg
Front Glass	High transmission tempered glass
Frame	Anodized aluminium alloy
Junction Box	IP67
Cable	4mm² (IEC)
Packaging Configuration	30pcs/box, 840pcs/40 HQ Container

Temperature Characteristics

Temperature Coefficient of Pmax	γ (Pm)	-0.39%/K
Temperature Coefficient of Voc	β (Voc)	-0.29%/K
Temperature Coefficient of Isc	α (Isc)	0.048%/K

Warranty

15yr 10-year product warranty
25yr 25-year warranty on power output

Added Value from Warranty

100% 97% 90% 80%

1 5 10 15 20 25 Year

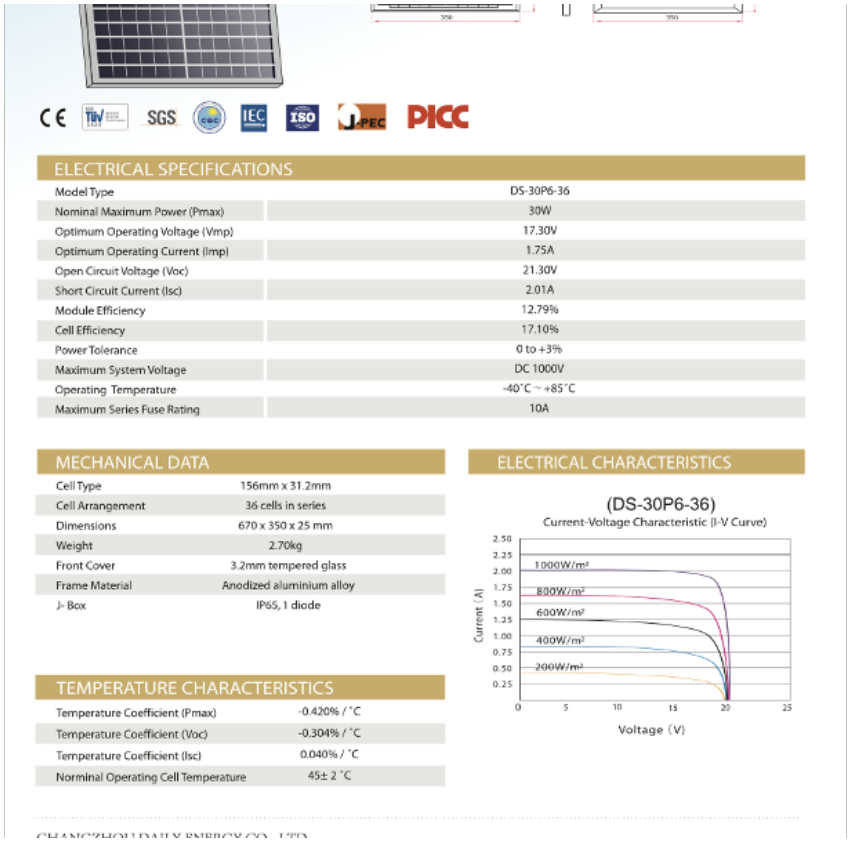
Information Box

Shanghai Aerospace Automobile Electromechanical Co., Ltd.
website: www.ht-saae.com
Email:Japan@ht-saae.com
Factory: Lianyungang ShenZhou New Energy Co., Ltd.

Figure: Data sheet design variation 1

Introduction: Challenges

Variability in Data Sheet Design



ELECTRICAL SPECIFICATIONS

Model Type	DS-30P6-36
Nominal Maximum Power (Pmax)	30W
Optimum Operating Voltage (Vmp)	17.30V
Optimum Operating Current (Imp)	1.75A
Open Circuit Voltage (Voc)	21.30V
Short Circuit Current (Isc)	2.01A
Module Efficiency	12.79%
Cell Efficiency	17.10%
Power Tolerance	0 to +3%
Maximum System Voltage	DC 1000V
Operating Temperature	-40˚C~+85˚C
Maximum Series Fuse Rating	10A

MECHANICAL DATA

Cell Type	156mm x 31.2mm
Cell Arrangement	36 cells in series
Dimensions	670 x 350 x 25 mm
Weight	2.70kg
Front Cover	3.2mm tempered glass
Frame Material	Anodized aluminium alloy
J- Box	IP65, 1 diode

TEMPERATURE CHARACTERISTICS

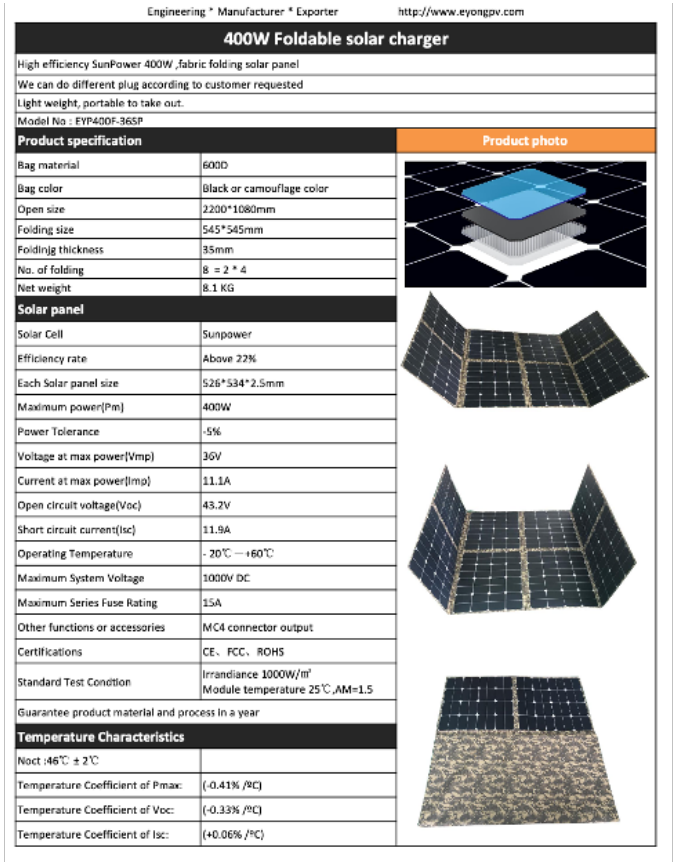
Temperature Coefficient (Pmax)	-0.420% / ˚C
Temperature Coefficient (Voc)	-0.304% / ˚C
Temperature Coefficient (Isc)	0.040% / ˚C
Norminal Operating Cell Temperature	45± 2 ˚C

ELECTRICAL CHARACTERISTICS

Figure: Data sheet design variation 2

Introduction: Challenges

Variability in Data Sheet Design



Engineering * Manufacturer * Exporter http://www.eyongpv.com

400W Foldable solar charger

High efficiency SunPower 400W ,fabric folding solar panel

We can do different plug according to customer requested

Light weight, portable to take out.

Model No : EYP400F-36SP

Product specification		Product photo
Bag material	600D	
Bag color	Black or camouflage color	
Open size	2200*1080mm	
Folding size	545*545mm	
Foldinjg thickness	35mm	
No. of folding	8 = 2 * 4	
Net weight	8.1 KG	
Solar panel		
Solar Cell	Sunpower	
Efficiency rate	Above 22%	
Each Solar panel size	526*534*2.5mm	
Maximum power(Pm)	400W	
Power Tolerance	-5%	
Voltage at max power(Vmp)	36V	
Current at max power(Imp)	11.1A	
Open circuit voltage(Voc)	43.2V	
Short circuit current(Isc)	11.9A	
Operating Temperature	- 20℃ — +60℃	
Maximum System Voltage	1000V DC	
Maximum Series Fuse Rating	15A	
Other functions or accessories	MC4 connector output	
Certifications	CE、FCC、ROHS	
Standard Test Condition	Irrandiance 1000W/m² Module temperature 25℃,AM=1.5	
Guarantee product material and process in a year		
Temperature Characteristics		
Noct :46℃ ± 2℃		
Temperature Coefficient of Pmax:	(-0.41% /ºC)	
Temperature Coefficient of Voc:	(-0.33% /ºC)	
Temperature Coefficient of Isc:	(+0.06% /ºC)	

Figure: Data sheet design variation 3

Introduction: Challenges

Variability in Data Sheet Design

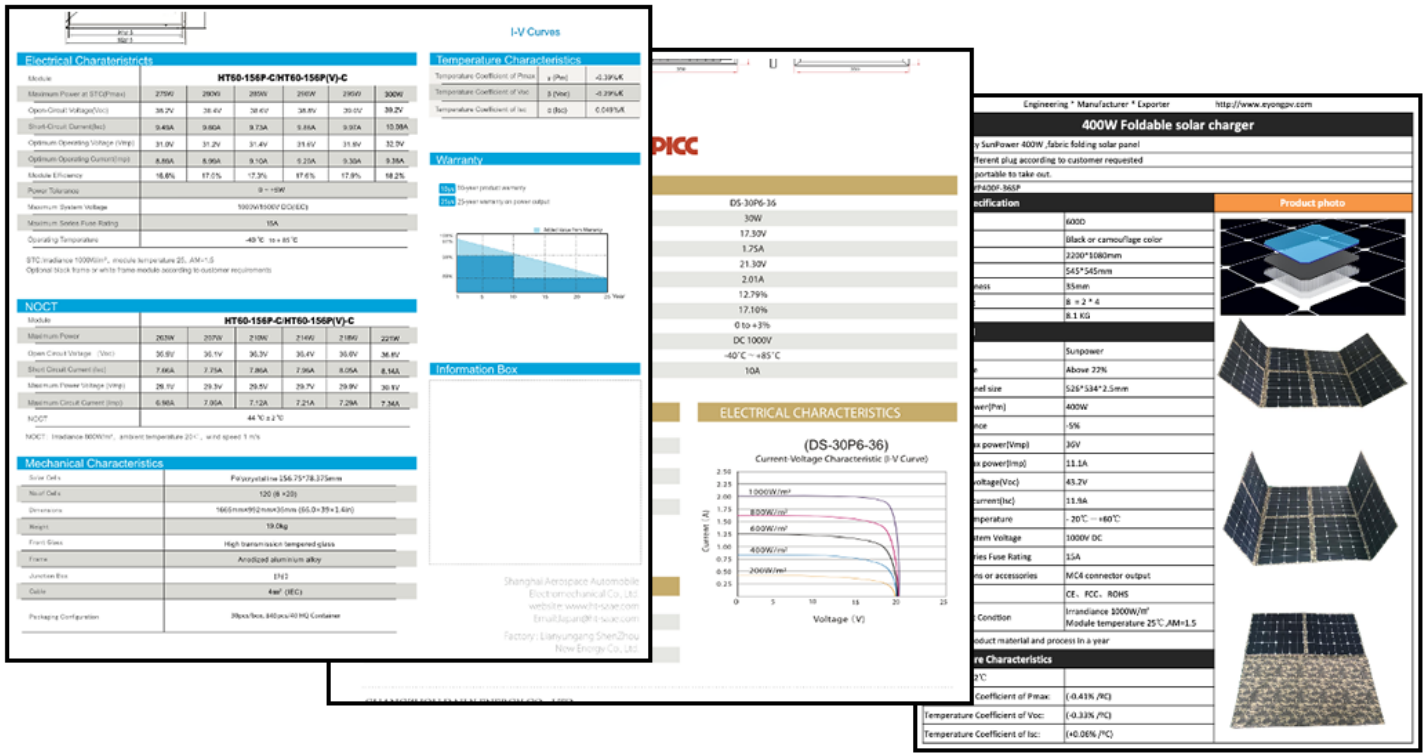


Figure: Data sheet design variations

Introduction: Existing Solutions

Shortcomings of Existing Off-the-Shelf Solutions

- Existing rule-based solutions
  - Tabula [1]
  - Camelot [2]
- Problems uncovered [3]
  - Failure in detection of tables
  - Overlap of detected table areas
  - Failure to extract required values

Introduction: Summary

- Existing off-the-shelf solutions fail on solar cell data sheets
- The variability in the data sheet design presents a challenge
- Manual labour not feasible for large quantities of documents
- **Goal:** An end-to-end solution for automatic extraction of solar cell data from tables in the PDF data sheets

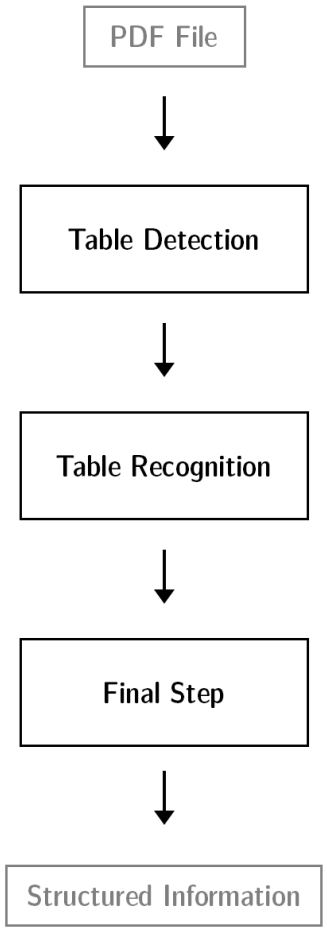
Introduction

Questions?

Approach

Approach: Overview

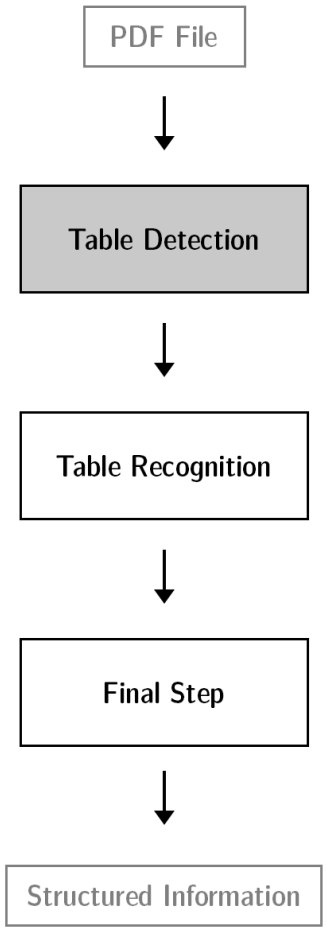
Major Steps Involved in Approach



1. **Table Detection**
   Locating where the tables are
2. **Table Recognition**
   Extracting the raw values from the tables
3. **Final Step**
   Validating and structuring extracted raw values

Approach: Overview

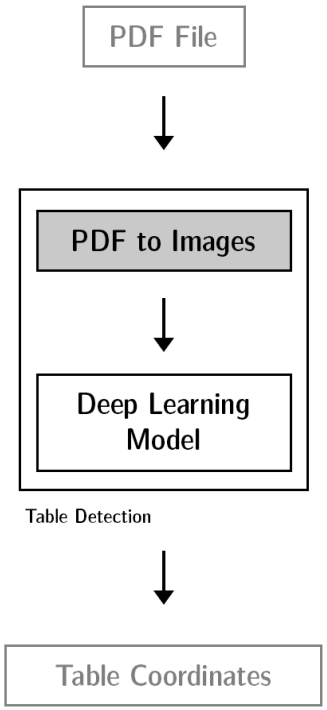
Major Steps Involved in Approach



1. **Table Detection**
   Locating where the tables are
2. **Table Recognition**
   Extracting the raw values from the tables
3. **Final Step**
   Validating and structuring extracted raw values

Approach: Table Detection

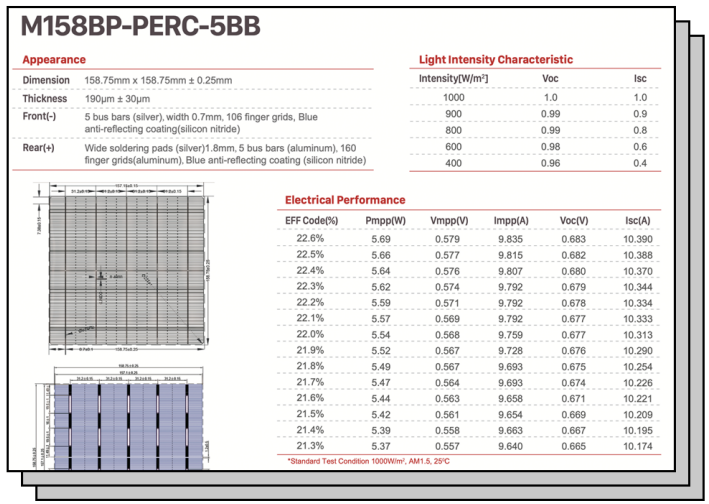
Locating tables in the documents



- Pages in PDF are converted to images

PDF Document

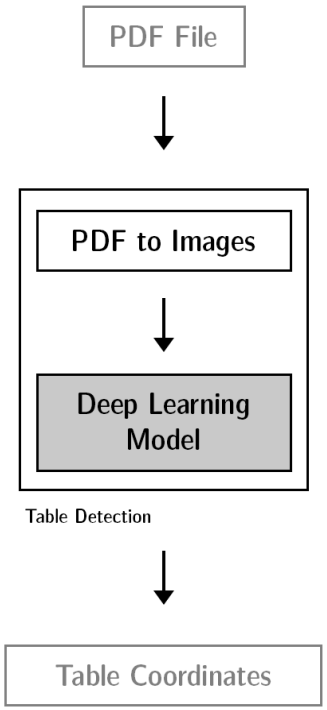
Convert Pages of PDF to Images



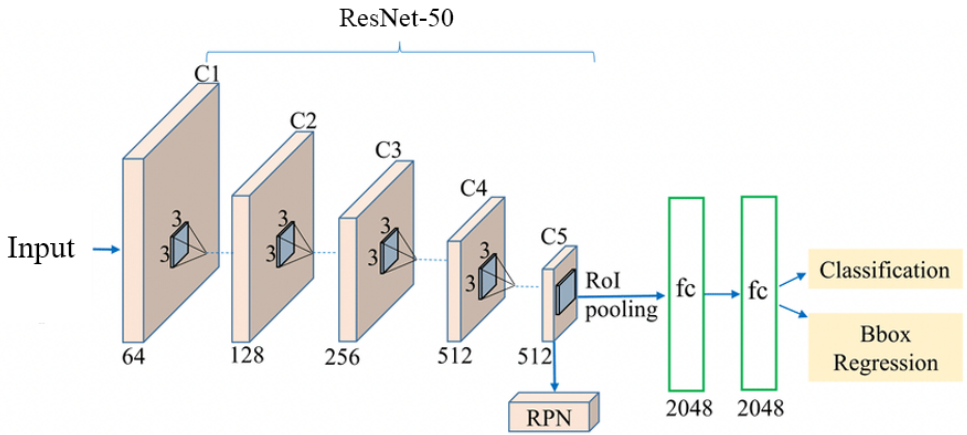
Images from PDF

Approach: Table Detection

Locating tables in the documents

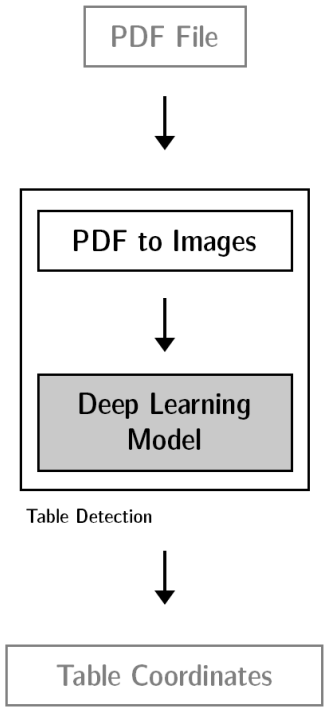


- An example Object Detector architecture (FasterRCNN) [4]

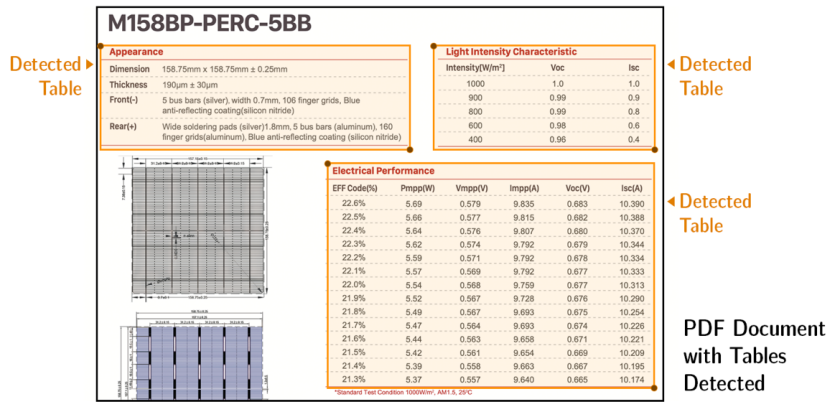


Approach: Table Detection

Locating tables in the documents



- Detected Tables are saved as CSV

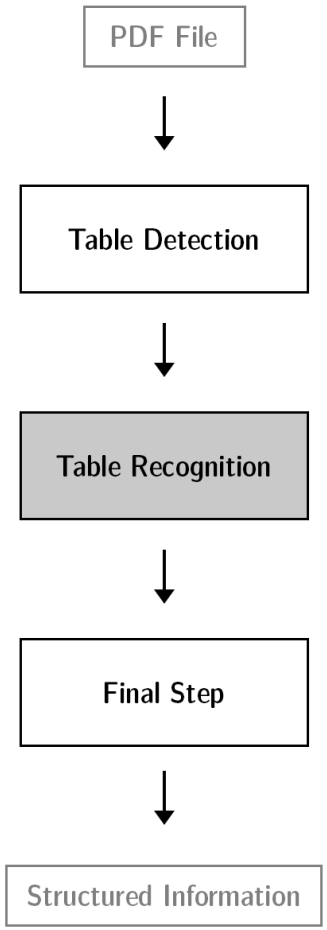



```
filename,pageno,x1,y1,x2,y2
1-4.pdf,1,252,289,575,215
1-4.pdf,1,254,530,576,325
1-4.pdf,1,359,647,576,556
1-4.pdf,1,50,237,212,190
1-4.pdf,1,38,647,335,557
```

CSV File containing Table Coordinates

Approach: Overview

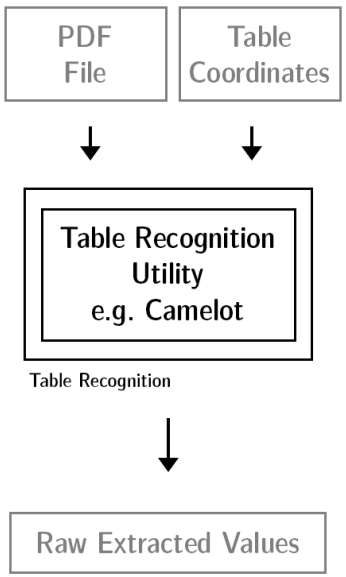
Major Steps Involved in Approach



1. **Table Detection**
   Locating where the tables are
2. **Table Recognition**
   Extracting the raw values from the tables
3. **Final Step**
   Validating and structuring extracted raw values

Approach: Table Recognition

Extracting raw values from table



PDF Document

+

```
filename,pageno,x1,y1,x2,y2
1-4.pdf,1,252,289,575,215
1-4.pdf,1,254,530,576,325
1-4.pdf,1,359,647,576,556
1-4.pdf,1,50,237,212,190
1-4.pdf,1,38,647,335,557
```

CSV File containing Table Coordinates

Extract Raw Table Values

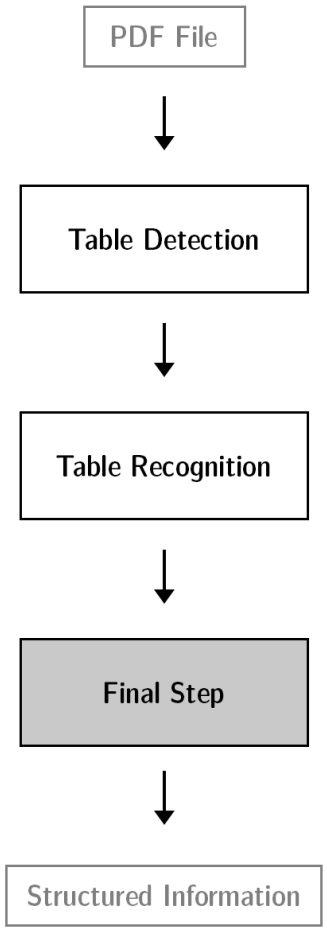
	A	B	C	D	E	F
1	EFF Code(%	Pmpp(W)	Vmpp(V)	Impp(A)	Voc(V)	Isc(A)
2	22.6%	5.69	0.579	9.835	0.683	10.390
3	22.5%	5.66	0.577	9.815	0.682	10.388
4	22.4%	5.64	0.576	9.807	0.680	10.370
5	22.3%	5.62	0.574	9.792	0.679	10.344
6	22.2%	5.59	0.571	9.792	0.678	10.334
7	22.1%	5.57	0.569	9.792	0.677	10.333
8	22.0%	5.54	0.568	9.759	0.677	10.313
9	21.9%	5.52	0.567	9.728	0.676	10.290
10	21.8%	5.49	0.567	9.693	0.675	10.254
11	21.7%	5.47	0.564	9.693	0.674	10.226
12	21.6%	5.44	0.563	9.658	0.671	10.221
13	21.5%	5.42	0.561	9.654	0.669	10.209
14	21.4%	5.39	0.558	9.663	0.667	10.195
15	21.3%	5.37	0.557	9.640	0.665	10.174

Raw Values Extracted from Tables

Figure: Raw values are extracted from tables

Approach: Overview

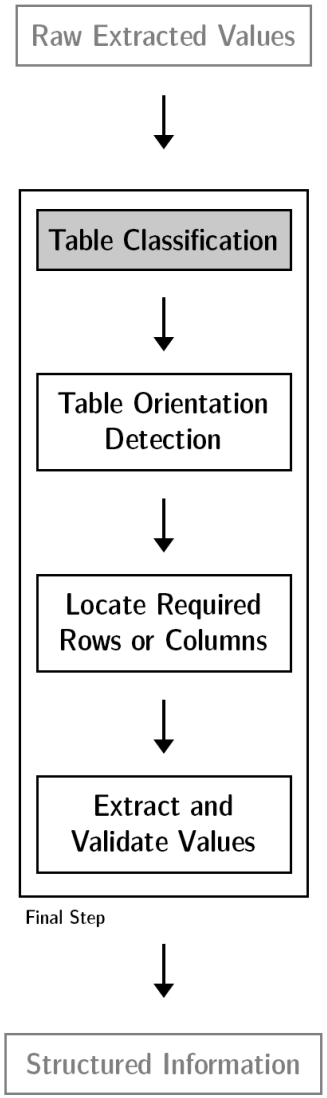
Major Steps Involved in Approach



1. **Table Detection**
   Locating where the tables are
2. **Table Recognition**
   Extracting the raw values from the tables
3. **Final Step**
   Validating and structuring extracted raw values

Approach: Final Step

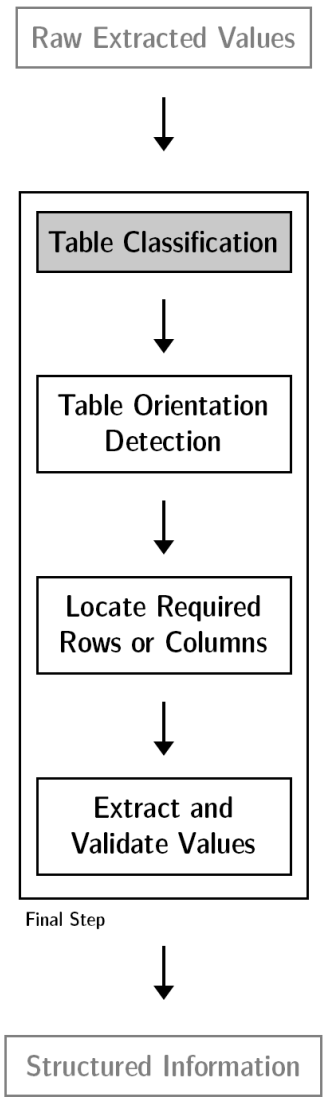
Table Classification



- Table type needs to be identified before the values can be validated
- The table title might be missing or not detected
- Can use the table content to identify the table type

Approach: Final Step

Table Classification



EFF Code(%)	Pmpp(W)	Vmpp(V)	Impp(A)	Voc(V)	Isc(A)
22.6%	5.69	0.579	9.835	0.683	10.390
22.5%	5.66	0.577	9.815	0.682	10.388
22.4%	5.64	0.576	9.807	0.680	10.370
22.3%	5.62	0.574	9.792	0.679	10.344
22.2%	5.59	0.571	9.792	0.678	10.334
22.1%	5.57	0.569	9.792	0.677	10.333
22.0%	5.54	0.568	9.759	0.677	10.313
21.9%	5.52	0.567	9.728	0.676	10.290
21.8%	5.49	0.567	9.693	0.675	10.254
21.7%	5.47	0.564	9.693	0.674	10.226
21.6%	5.44	0.563	9.658	0.671	10.221
21.5%	5.42	0.561	9.654	0.669	10.209
21.4%	5.39	0.558	9.663	0.667	10.195
21.3%	5.37	0.557	9.640	0.665	10.174

Extract Keywords

[`EFF`, `Code`, `%`, `Pmpp`, `W`, `Vmpp`, `V`, `Impp`, `A`, ...]

Vectorise Keywords

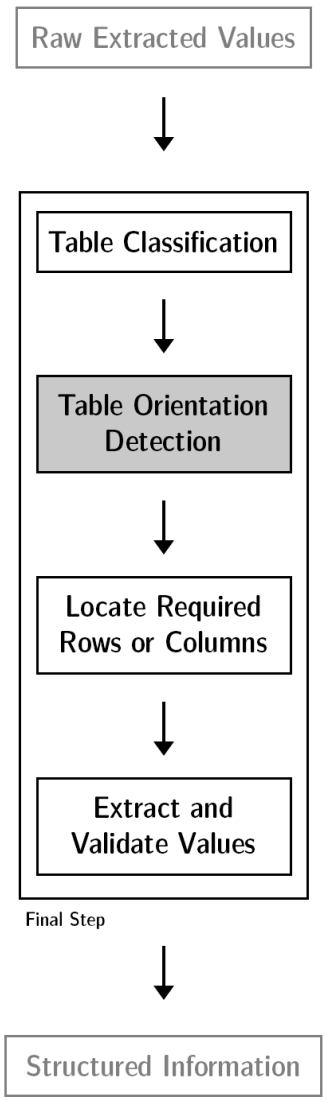
[`0`, `1`, `0`, `0`, `0`, `1`, `0`, `0`, `1`, ...]

Classify Table

Electrical Characteristics

Approach: Final Step

Table Orientation Detection



Horizontal Table

Efficiency Code		196	195	194	193	192	191	190	189
Efficiency	Eff(%)	19.60	19.50	19.40	19.30	19.20	19.10	19.00	18.90
Power	Ppm(W)	4.68	4.66	4.64	4.61	4.59	4.56	4.54	4.52
Max. Power Current	Ipm(A)	8.64	8.61	8.58	8.55	8.53	8.51	8.49	8.48
Short Circuit Current	Isc(A)	9.14	9.11	9.08	9.05	9.03	9.02	9.01	9.01
Max. Power Voltage	Vpm(V)	0.542	0.541	0.541	0.539	0.538	0.536	0.535	0.533
Open Circuit Voltage	Voc(V)	0.643	0.643	0.642	0.641	0.641	0.640	0.639	0.638

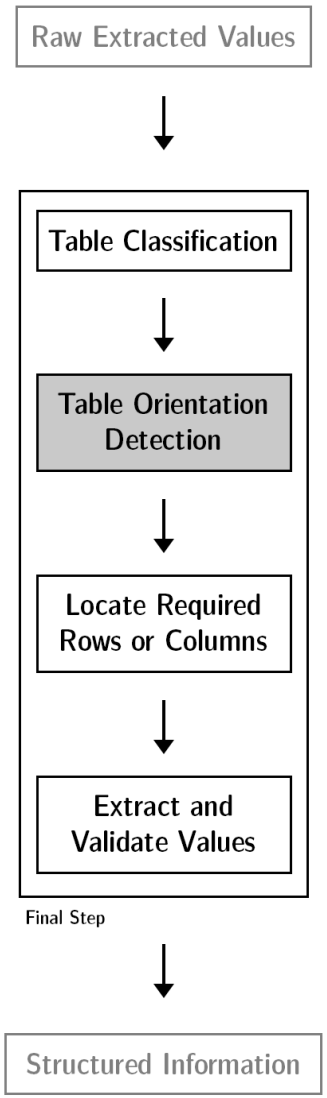
Vertical Table

ELECTRICAL PARAMETERS					
Efficiency Ncell(%)	Maximum power Pmpp(w)	Peak voltage Vmp(V)	Peak current Imp(A)	Open-circuit voltage Voc(V)	Short-circuit current Isc(A)
⩾18.6	4.53	0.540	8.389	0.641	8.889
18.4-18.6	4.48	0.539	8.311	0.641	8.802
18.2-18.4	4.43	0.538	8.234	0.640	8.729
18.0-18.2	4.38	0.536	8.172	0.639	8.655
17.8-18.0	4.33	0.534	8.109	0.638	8.580
17.6-17.8	4.28	0.531	8.060	0.636	8.518
17.4-17.6	4.23	0.528	8.011	0.634	8.456

- Table can be vertical or horizontal
- Need to detect before the next steps

Approach: Final Step

Table Orientation Detection



Eff （%）	17.9	17.8	17.7	17.6	17.5	17.4	17.3	17.2
Pmpp （W）	4.359	4.334	4.308	4.283	4.259	4.234	4.210	4.186
Uoc(V)	0.630	0.629	0.629	0.628	0.627	0.626	0.626	0.624
Isc(A)	8.760	8.747	8.709	8.680	8.656	8.648	8.626	8.573
Umpp(V)	0.530	0.529	0.527	0.524	0.524	0.524	0.523	0.520
Impp(A)	8.233	8.194	8.168	8.173	8.133	8.088	8.053	8.052
FF(%)	78.973	78.785	78.639	78.567	78.441	78.240	77.970	78.290

Extract Values

17.9	17.8	17.7	17.6	17.5	17.4	17.3	17.2
4.359	4.334	4.308	4.283	4.259	4.234	4.210	4.186
0.630	0.629	0.629	0.628	0.627	0.626	0.626	0.624
8.760	8.747	8.709	8.680	8.656	8.648	8.626	8.573
0.530	0.529	0.527	0.524	0.524	0.524	0.523	0.520
8.233	8.194	8.168	8.173	8.133	8.088	8.053	8.052
78.973	78.785	78.639	78.567	78.441	78.240	77.970	78.290

Calculate Variances

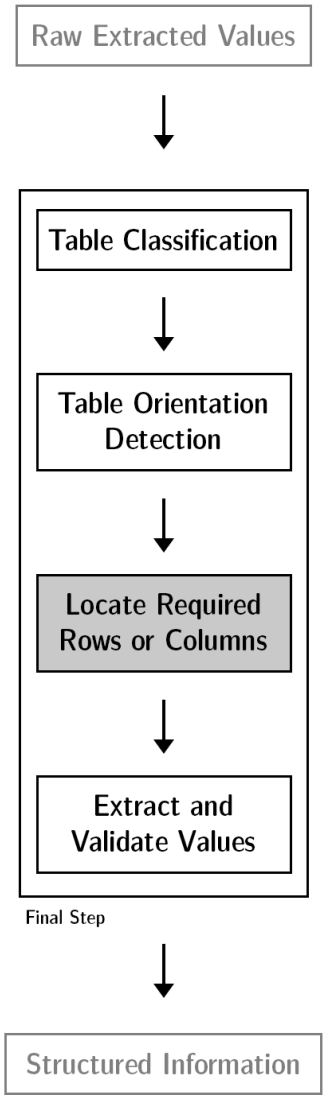
$$\text{row-variance} = 0.1755$$
$$\text{column-variance} = 6173.95$$

Decide Orientation

Horizontal Table

Approach: Final Step

Locating Required Rows or Columns



- Locate the required rows or columns using Regex pattern matching

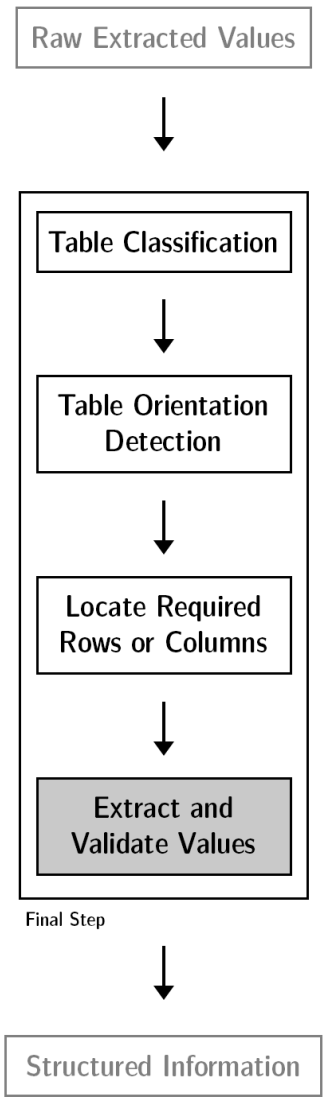
Eff (%)	17.9	17.8	17.7	17.6	17.5	17.4	17.3	17.2
Pmpp (W)	4.359	4.334	4.308	4.283	4.259	4.234	4.210	4.186
Uoc(V)	0.630	0.629	0.629	0.628	0.627	0.626	0.626	0.624
Isc(A)	8.760	8.747	8.709	8.680	8.656	8.648	8.626	8.573
Umpp(V)	0.530	0.529	0.527	0.524	0.524	0.524	0.523	0.520
Impp(A)	8.233	8.194	8.168	8.173	8.133	8.088	8.053	8.052
FF(%)	78.973	78.785	78.639	78.567	78.441	78.240	77.970	78.290

Locate Open Circuit Voltage Row

Uoc(V)	0.630	0.629	0.629	0.628	0.627	0.626	0.626	0.624

Approach: Final Step

Validating and structuring raw values



- Extract the valid values from the located row

Uoc(V)	0.630	0.629	0.629	0.628	0.627	0.626	0.626	0.624

Extract Values

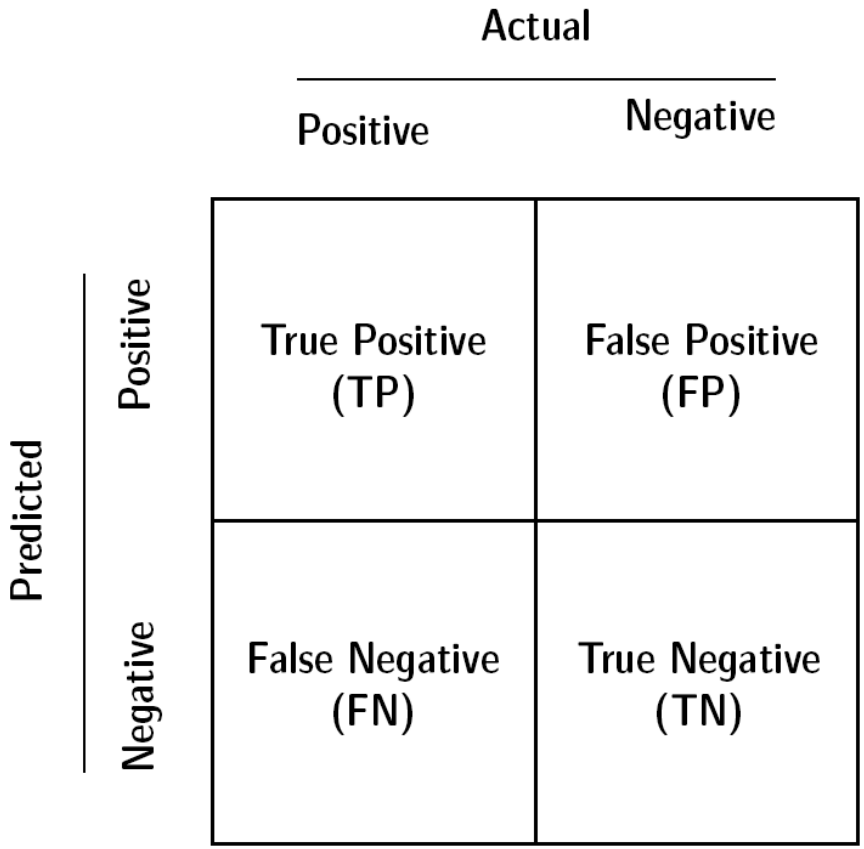
Open Circuit Voltage = [`0.63`, `0.629`, `0.629`, `0.628`, `0.627`, ...]

Approach

Questions?

Experiments

Experiments: General Evaluation Metrics



- Confusion matrix values for each class will be calculated

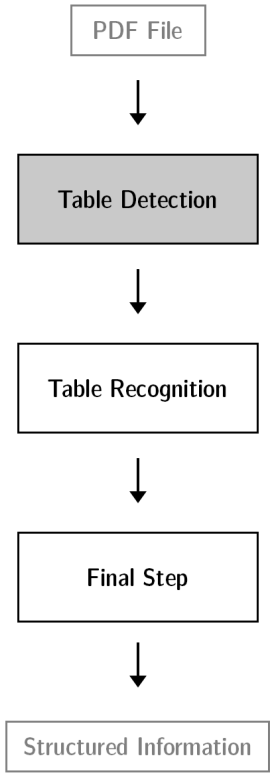
Experiments: General Evaluation Metrics

- Confusion matrix values for each class will be calculated
  - **TP**: Actually positive, detected as positive
  - **FP**: Actually negative, detected as positive
  - **FN**: Actually positive, detected as negative
- Precision = $\frac{TP}{TP+FP}$
- Recall = $\frac{TP}{TP+FN}$
- F1-Score = $\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$

Experiments: Table Detection Evaluation

Overview

This experiment evaluates the approach described for the table detection step.



Experiments: Table Detection Evaluation

Evaluation Metrics

- IoU Overlap Threshold
  - 75% IoU Overlap Threshold
  - 90% IoU Overlap Threshold

Experiments: Table Detection Evaluation

Evaluation Metrics

- IoU Overlap Threshold
  - 75% IoU Overlap Threshold
  - 90% IoU Overlap Threshold
- Precision = $\frac{TP}{TP+FP}$
  How many guesses made by model were actually tables?
- Recall = $\frac{TP}{TP+FN}$
  How many actual tables were detected by the model?

Experiments: Table Detection Evaluation

Evaluation Metrics

- IoU Overlap Threshold
  - 75% IoU Overlap Threshold
  - 90% IoU Overlap Threshold
- Precision = $\frac{TP}{TP+FP}$
  How many guesses made by model were actually tables?
- Recall = $\frac{TP}{TP+FN}$
  How many actual tables were detected by the model?

Note

For selecting the best model Recall performance at 90% IoU threshold was considered

Experiments: Table Detection Evaluation

Dataset

- Randomly selected PDF documents
- Images: 2675 in total
- Tables: 5896 in total
- The tables were manually labelled

Experiments: Table Detection Evaluation

Setup

- Models
  - Single Stage Detectors
    - RetinaNet [5]
    - RetinaNet v2 [5]
  - Two Stage Detectors
    - FasterRCNN [6]
    - FasterRCNN v2 [6]

Experiments: Table Detection Evaluation

Setup

- Models
  - Single Stage Detectors
    - RetinaNet [5]
    - RetinaNet v2 [5]
  - Two Stage Detectors
    - FasterRCNN [6]
    - FasterRCNN v2 [6]
- Method
  - Trained for 50 epochs
  - Trained on NVidia P100 GPUs on Kaggle
  - Initial weights: Microsoft COCO [7]
  - Hyper-parameter Tuning
    - Batch Size
    - Learning Rate
  - Training-Evaluation Split
    - 80% for training
    - 20% for evaluation

Experiments: Table Detection Evaluation

Results

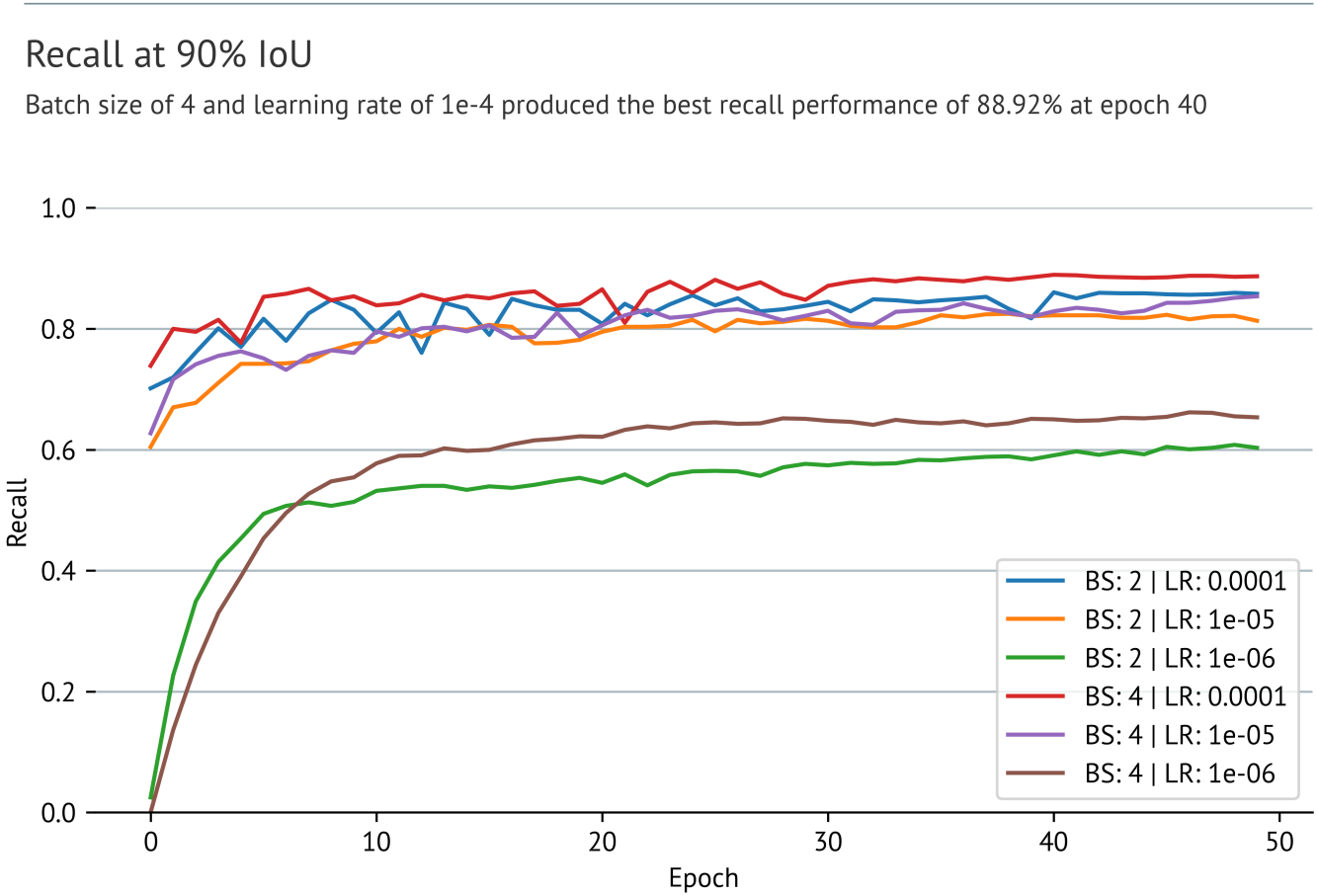


Figure: FasterRCNN V2 Hyperparameter Tuning

Experiments: Table Detection Evaluation

Results

Architecture	Configuration		Epoch	Recall		Precision	
	BS	LR		90% IoU	75% IoU	90% IoU	75% IoU
RetinaNet	4	1e-04	47	0.8007	0.9247	0.8155	0.9419
RetinaNet v2	4	1e-04	34	0.8007	0.9280	0.8259	0.9573
FasterRCNN	2	1e-04	23	0.8817	0.9520	0.8825	0.9528
FasterRCNN v2	**4**	**1e-04**	**40**	**0.8892**	**0.9512**	**0.9018**	**0.9648**

Table: Object Detector Model Comparison

- For the required task:
  - Two-stage detectors perform better than single-stage detectors
  - FasterRCNN v2 performs the best

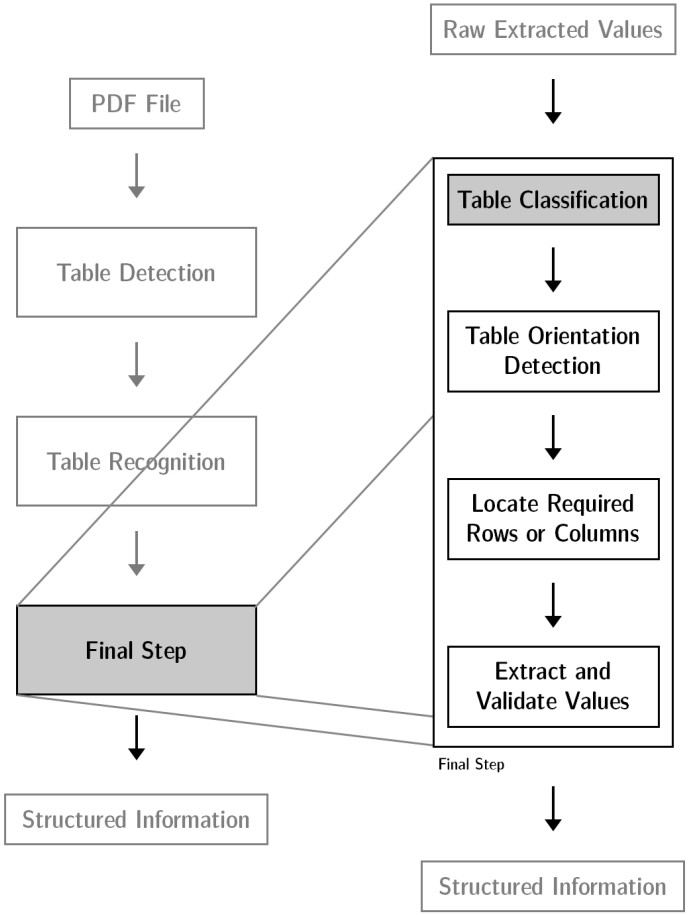
Experiments: Table Detection Evaluation

Questions?

Experiments: Table Classification Evaluation

Overview

This experiment evaluates the approach described for table classification as part of the final step.



Experiments: Table Classification Evaluation

Evaluation Metrics

- Precision = $\frac{TP}{TP+FP}$
  How many guesses made by model belonged to the correct class?
- Recall = $\frac{TP}{TP+FN}$
  Per class, how many actual tables were identified by the model?
- F1-Score = $\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

Note

F1-Score was used to determine the best table classification approach

Experiments: Table Classification Evaluation

Dataset

- Dataset
  - 60 excel files
  - 215 tables
- Classes
  - Electrical Characteristics (EC)
  - Thermal Characteristics (TC)
  - Mechanical Characteristics (MC)
  - Other (O)

Experiments: Table Classification Evaluation

Setup

- Word vectorisers
  - Count Vectoriser [8]
  - TF-IDF [9]
- Classifiers
  - K Nearest Neighbours [10]
  - Naive Bayes [11]
- Cross compared using 5-Fold Cross Validation

Experiments: Table Classification Evaluation

Results

Approach	F1-Score for Classes			
	EC	TC	MC	O
K Nearest Neighbours & Count Vectoriser	0.89	0.81	0.89	0.83
K Nearest Neighbours & TF-IDF	0.72	0.87	0.77	0.74
Naive Bayes & Count Vectoriser	0.95	0.80	0.96	0.85
Naive Bayes & TF-IDF	**0.95**	**0.95**	**0.94**	**0.93**

Table: **Comparison of Table Classification Approaches** The F1-Scores for each class are averaged over the 5-fold cross validation run

Experiments: Table Classification Evaluation

Results

Approach	F1-Score for Classes			
	EC	TC	MC	O
K Nearest Neighbours & Count Vectoriser	0.89	0.81	0.89	0.83
K Nearest Neighbours & TF-IDF	0.72	0.87	0.77	0.74
Naive Bayes & Count Vectoriser	0.95	0.80	0.96	0.85
Naive Bayes & TF-IDF	**0.95**	**0.95**	**0.94**	**0.93**

Table: **Comparison of Table Classification Approaches** The F1-Scores for each class are averaged over the 5-fold cross validation run

- For the required task:
  - Naive Bayes is a better classifier than K Nearest Neighbours
  - The combination of Naive Bayes with TF-IDF performs the best

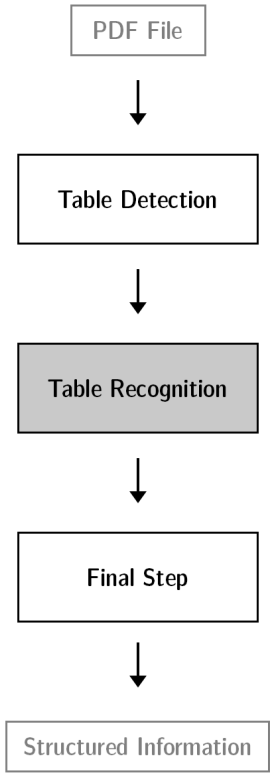
Experiments: Table Classification Evaluations

Questions?

Experiments: Complete Pipeline Evaluation

Overview

This experiment will evaluate the complete pipeline as well as the table recognition utilities.



Experiments: Complete Pipeline Evaluation

Evaluation Metrics

- Precision = $\frac{TP}{TP+FP}$
  How many of the values extracted were the actual values?
- Recall = $\frac{TP}{TP+FN}$
  How many of the actual values were extracted?
- F1-Score = $\frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$

Note

F1-Score was used to compare table recognition utilities

Experiments: Complete Pipeline Evaluation

Dataset

- 10 PDF documents
- Diverse selection
- Ground-truth
  - Electrical Characteristics
  - Thermal Characteristics
  - Manually extracted and saved in a YAML file

Experiments: Complete Pipeline Evaluation

Table Recognition Approaches Compared

- **Baseline**
  Custom utility created using raw text and positional information from the PDF.
- **Tabula** [1]
  Open-source Python utility for extracting tabular data from PDF documents.
- **Camelot** [2]
  Open-source Python utility for extracting tabular data from PDF documents.

Experiments: Complete Pipeline Evaluation

Setup

1. 10 PDFs are fed into the pipeline and electrical and thermal characteristics are extracted
2. The extracted values are compared against ground truth values and precision, recall and f1-scores are calculated.
3. The process is repeated for all the table recognition utilities.

Experiments: Complete Pipeline Evaluation

Results

Approach	Precision	Recall	F1-Score
Baseline	0.7714	0.7066	0.7376
Camelot	**0.9986**	**0.9490**	**0.9731**
Tabula	0.8405	0.7769	0.8074

Table: Effect of different table recognition approaches on performance of complete pipeline

Experiments: Complete Pipeline Evaluation

Questions?

Conclusion

Key Takeaways

- Deep Learning based Object Detectors are well suited for Table Detection.
- Text Classification is a suitable technique for table type identification.
- Rule-based Table Recognition utilities are good enough for simple tables.
- The complete pipeline is well suited for extracting tabular information from solar cell data sheets and it would be interesting to test its feasibility on other domains as well.

Thank You!

Questions?

References I

[1] Tabula, "Tabula: Extract Tables from PDFs." `https://tabula.technology`, accessed on 2023-04-11.

[2] Camelot, "Camelot: PDF Table Extraction for Humans." `https://camelot-py.readthedocs.io/en/master/`, accessed on 2023-04-11.

[3] B. D. D. Artha, "Data mining im bereich von photovoltaik-technologie," 2021.

[4] R. Ding, L. Dai, G. Li, and H. Liu, "Tdd-net: a tiny defect detection network for printed circuit boards," *CAAI Transactions on Intelligence Technology*, vol. 4, no. 2, pp. 110–116, 2019.

[5] PyTorch, "RetinaNet." `https://pytorch.org/vision/main/models/retinanet.html`, accessed on 2023-04-10.

References II

[6] PyTorch, "FasterRCNN." `https://pytorch.org/vision/master/models/faster_rcnn.html`, accessed on 2023-04-10.

[7] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[8] Sci-Kit-Learn, "CountVectorizer." `https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html`, accessed on 2023-04-11.

References III

[9] Sci-Kit-Learn, "TfidfVectorizer." https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html, accessed on 2023-04-11.

[10] Sci-Kit-Learn, "Nearest Neighbors." https://scikit-learn.org/stable/modules/neighbors.html, accessed on 2023-04-11.

[11] Sci-Kit-Learn, "Naive Bayes." https://scikit-learn.org/stable/modules/naive_bayes.html, accessed on 2023-04-11.

Appendix: RetinaNet Training Run Graph

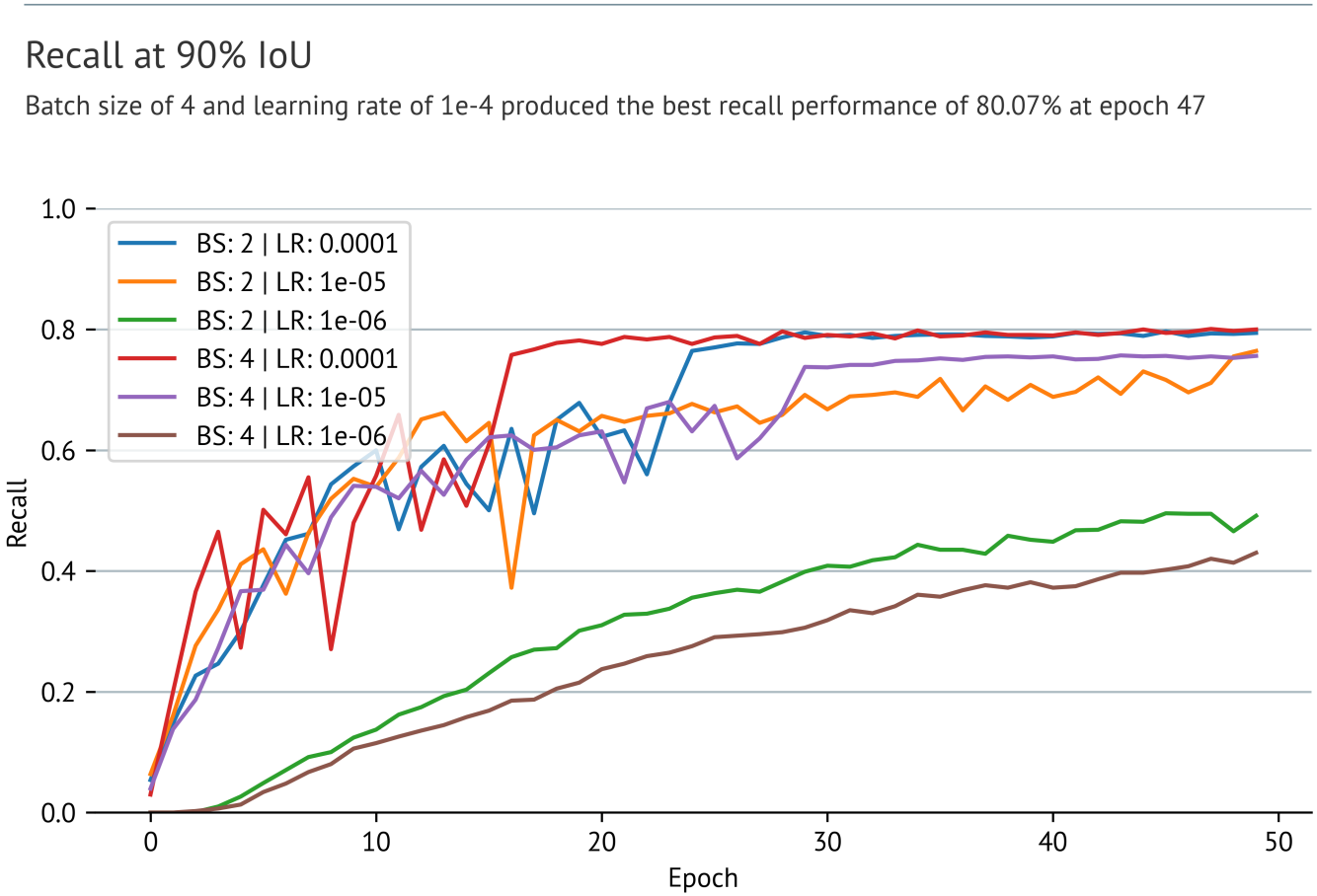


Figure: RetinaNet Hyperparameter Tuning

Appendix: RetinaNet Training Run Results

Configuration		Epoch	Recall		Precision	
Batch Size	Learning Rate		90% IoU	75% IoU	90% IoU	75% IoU
2	1e-04	45	0.7965	0.9322	0.8099	0.9479
2	1e-05	49	0.7643	0.9388	0.7732	0.9498
2	1e-06	45	0.4955	0.8859	0.5008	0.8955
4	**1e-04**	**47**	**0.8007**	**0.9247**	**0.8155**	**0.9419**
4	1e-05	43	0.7568	0.9247	0.7663	0.9363
4	1e-06	49	0.4301	0.8354	0.4758	0.9241

Table: **RetinaNet Results**

Appendix: RetinaNet v2 Training Run Graph

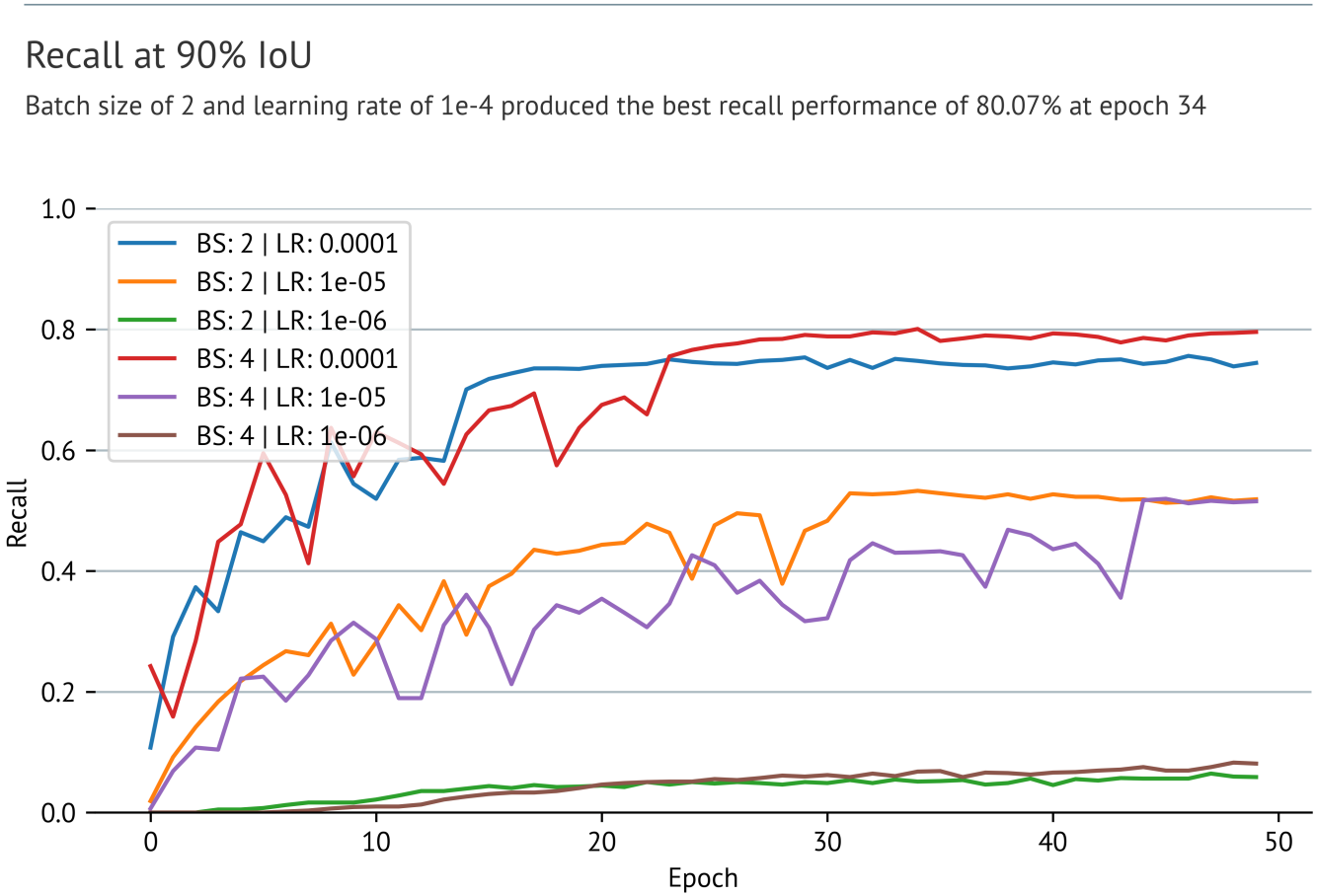


Figure: RetinaNet V2 Hyperparameter Tuning

Appendix: RetinaNet v2 Training Run Results

Configuration		Epoch	Recall		Precision	
Batch Size	Learning Rate		90% IoU	75% IoU	90% IoU	75% IoU
2	1e-04	46	0.756	0.9156	0.7455	0.9029
2	1e-05	34	0.5327	0.8519	0.5389	0.8619
2	1e-06	47	0.0645	0.5285	0.0707	0.5788
4	**1e-04**	**34**	**0.8007**	**0.928**	**0.8259**	**0.9573**
4	1e-05	45	0.5194	0.8594	0.5182	0.8573
4	1e-06	48	0.0827	0.5476	0.104	0.6881

Table: **RetinaNet V2 Results**

Appendix: FasterRCNN Training Run Graph

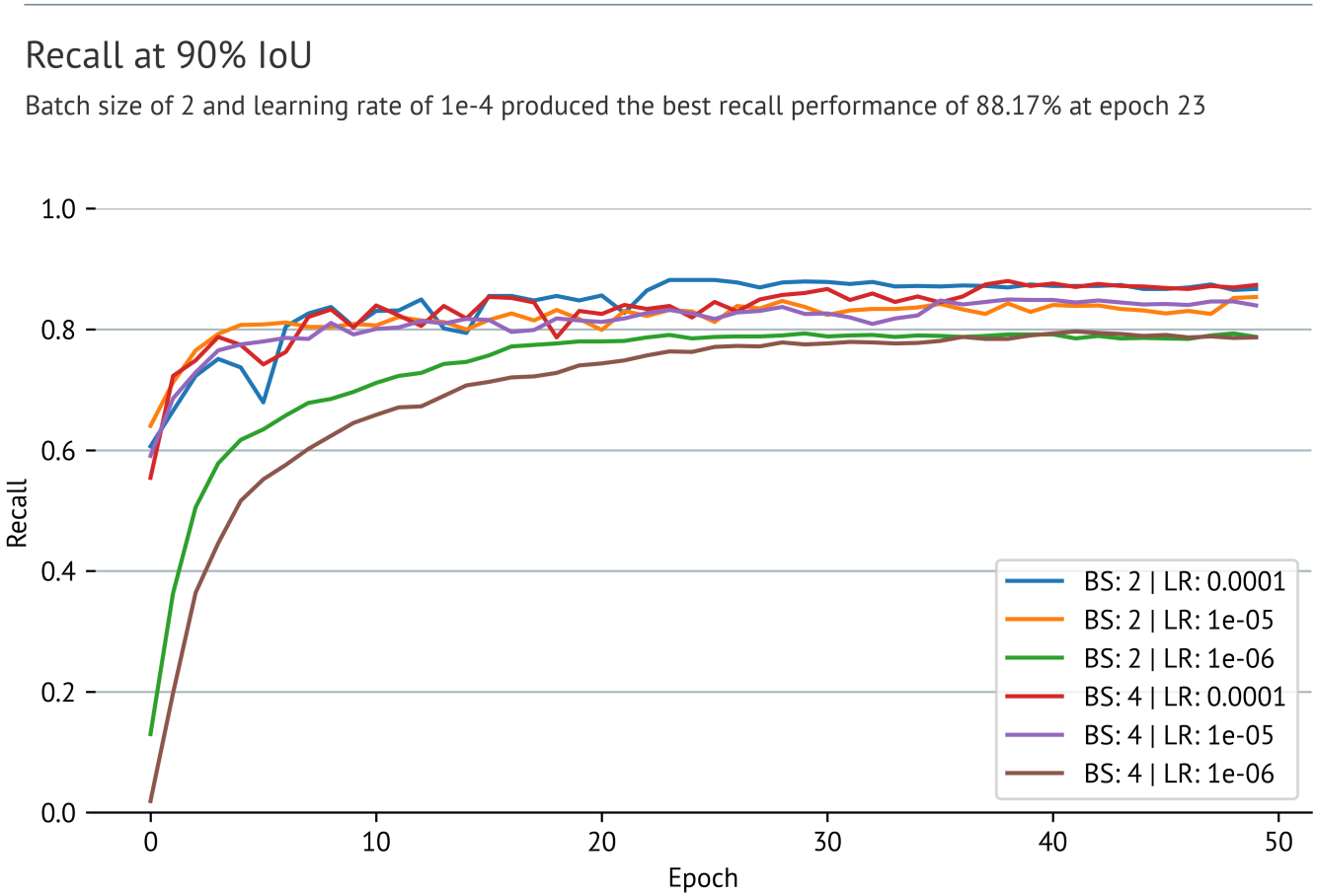


Figure: FasterRCNN Hyperparameter Tuning

Appendix: FasterRCNN Training Run Results

Configuration		Epoch	Recall		Precision	
Batch Size	Learning Rate		90% IoU	75% IoU	90% IoU	75% IoU
2	**1e-04**	**23**	**0.8817**	**0.952**	**0.8825**	**0.9528**
2	1e-05	49	0.8536	0.9545	0.8411	0.9405
2	1e-06	29	0.7932	0.957	0.7551	0.911
4	1e-04	38	0.8801	0.9495	0.8837	0.9535
4	1e-05	38	0.8495	0.9529	0.8397	0.9419
4	1e-06	41	0.7965	0.9504	0.7631	0.9105

Table: **FasterRCNN Results**

Appendix: FasterRCNN v2 Training Run Graph

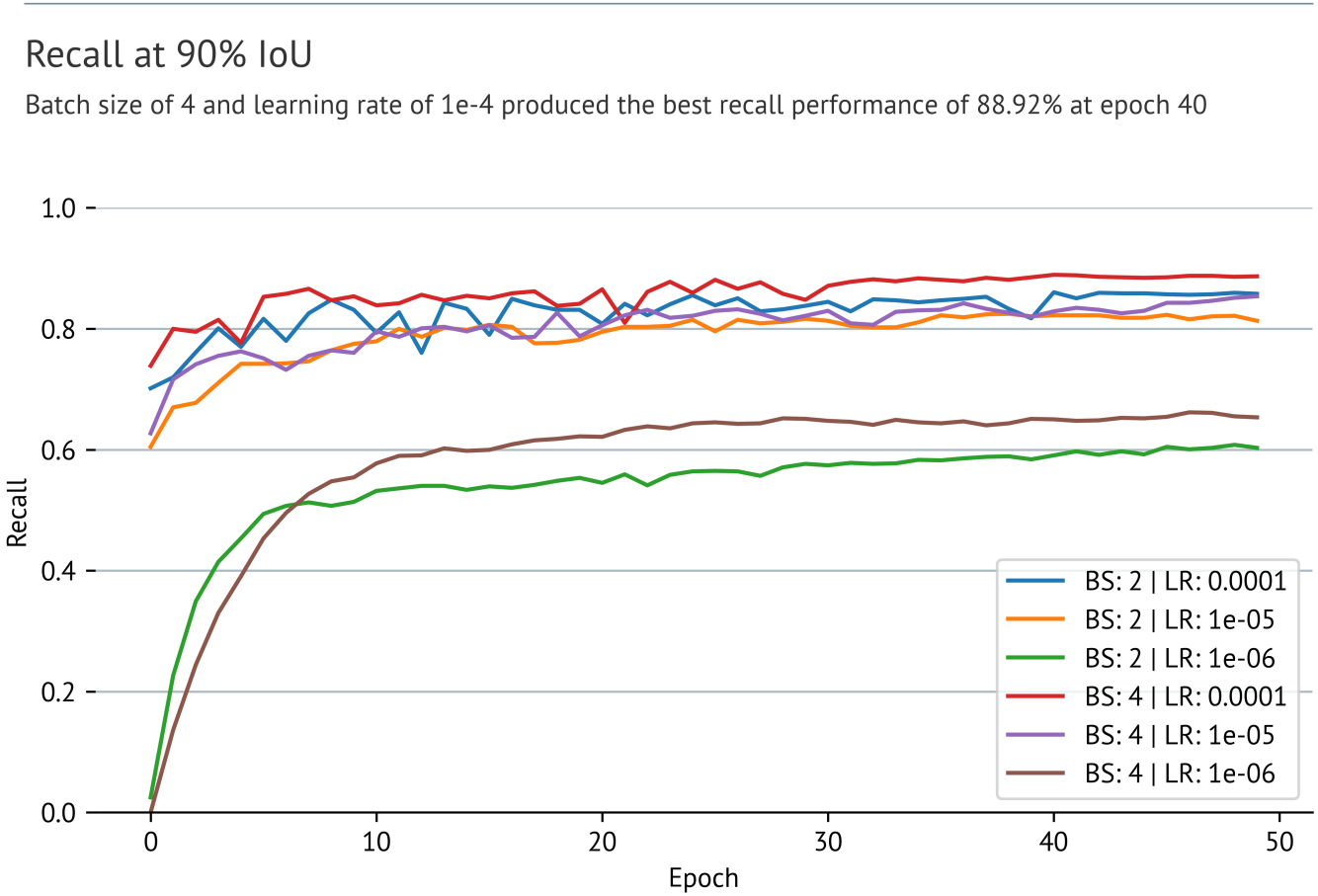


Figure: FasterRCNN V2 Hyperparameter Tuning

Appendix: FasterRCNN v2 Training Run Results

Configuration		Epoch	Recall		Precision	
Batch Size	Learning Rate		90% IoU	75% IoU	90% IoU	75% IoU
2	1e-04	40	0.8602	0.9322	0.8784	0.9519
2	1e-05	38	0.8246	0.9313	0.8295	0.9368
2	1e-06	48	0.6079	0.8759	0.5829	0.8398
4	**1e-04**	**40**	**0.8892**	**0.9512**	**0.9018**	**0.9648**
4	1e-05	49	0.8536	0.9429	0.8445	0.9329
4	1e-06	46	0.6617	0.8941	0.6385	0.8627

Table: **FasterRCNN V2 Results**

Appendix: Table Class - KNN with CV

Class	Precision	Recall	F1-Score
Electrical Characteristics	0.92	0.87	0.89
Thermal Characteristics	0.97	0.71	0.81
Mechanical Characteristics	0.95	0.87	0.89
Other	0.75	0.93	0.83
Overall Accuracy			0.86

Table: **K-Nearest Neighbours with Count Vectoriser** The table shows the averaged evaluation results of 5 Fold Cross Validation performed on the dataset.

Appendix: Table Class - KNN with TF-IDF

Class	Precision	Recall	F1-Score
Electrical Characteristics	0.95	0.6	0.72
Thermal Characteristics	1.00	0.78	0.87
Mechanical Characteristics	1.00	0.65	0.77
Other	0.60	0.98	0.74
Overall Accuracy			0.77

Table: **K-Nearest Neighbours with TF-IDF Vectoriser** The table shows the averaged evaluation results of 5 Fold Cross Validation performed on the dataset.

Appendix: Table Class - Naive Bayes with CV

Class	Precision	Recall	F1-Score
Electrical Characteristics	0.93	0.96	0.95
Thermal Characteristics	0.78	0.91	0.80
Mechanical Characteristics	0.93	1.0	0.96
Other	0.96	0.80	0.85
Overall Accuracy			0.89

Table: **Naive Bayes with Count Vectoriser** The table shows the averaged evaluation results of 5 Fold Cross Validation performed on the dataset.

Appendix: Table Class - Naive Bayes with TF-IDF

Class	Precision	Recall	F1-Score
Electrical Characteristics	0.92	0.98	0.95
Thermal Characteristics	1.00	0.91	0.95
Mechanical Characteristics	0.93	0.97	0.94
Other	0.94	0.92	0.93
Overall Accuracy			0.94

Table: **Naive Bayes with TF-IDF Vectoriser** The table shows the averaged evaluation results of 5 Fold Cross Validation performed on the dataset.