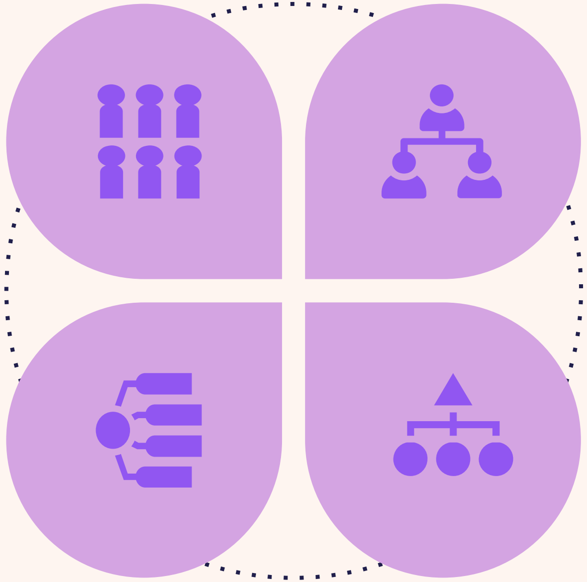
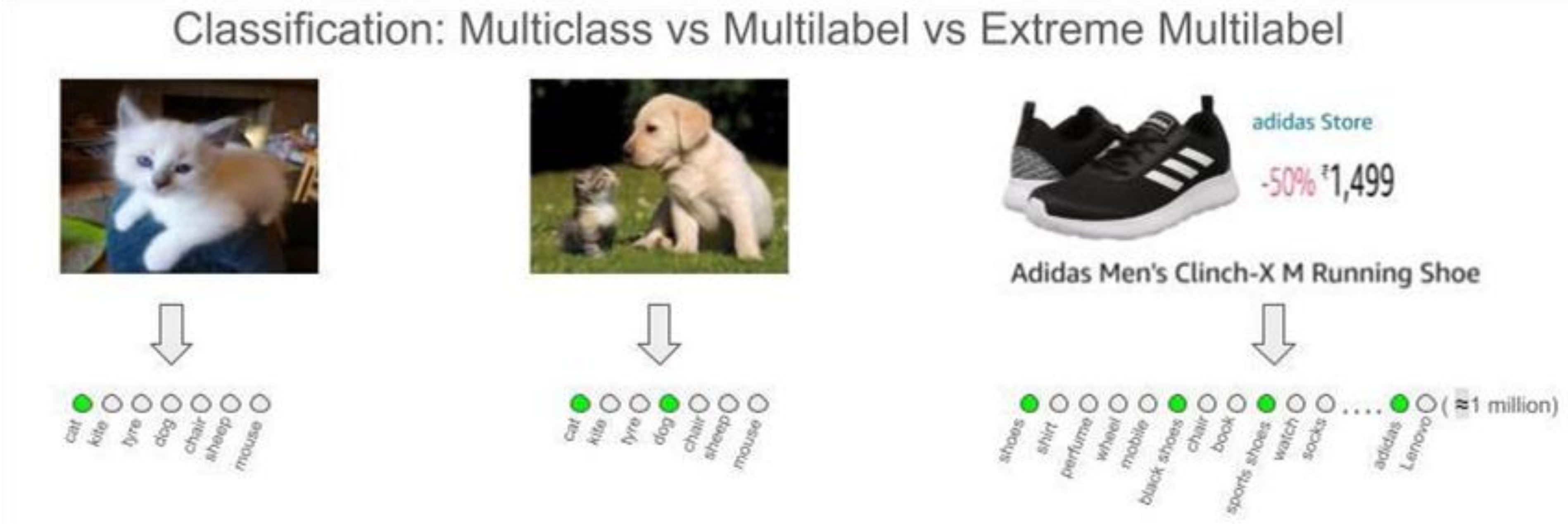
Extreme Multi Label Classification



Mahim Jain (20AE3AI02) | MTP-1

What is XMC?



Classification: Multiclass vs Multilabel vs Extreme Multilabel
cat kite tyre dog chair sheep mouse
cat kite tyre dog chair sheep mouse
adidas Store
-50% ₹1,499
Adidas Men's Clinch-X M Running Shoe
shoes shirt perfume wheel mobile black shoes chair book sports shoes watch socks .... adidas Lenovo (≈1 million)

Motivation for XMC

Content Retrieval

(e.g., Wikipedia)

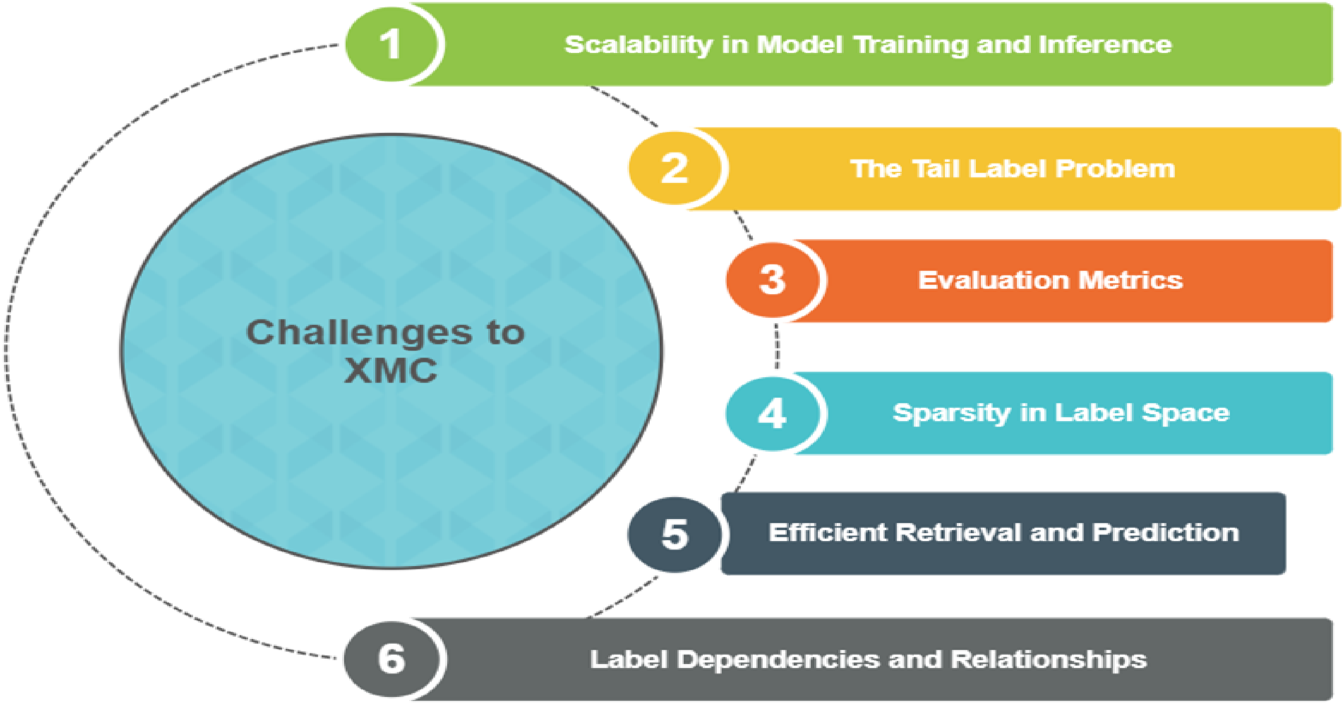
Recommendation Systems

(e.g., Amazon)

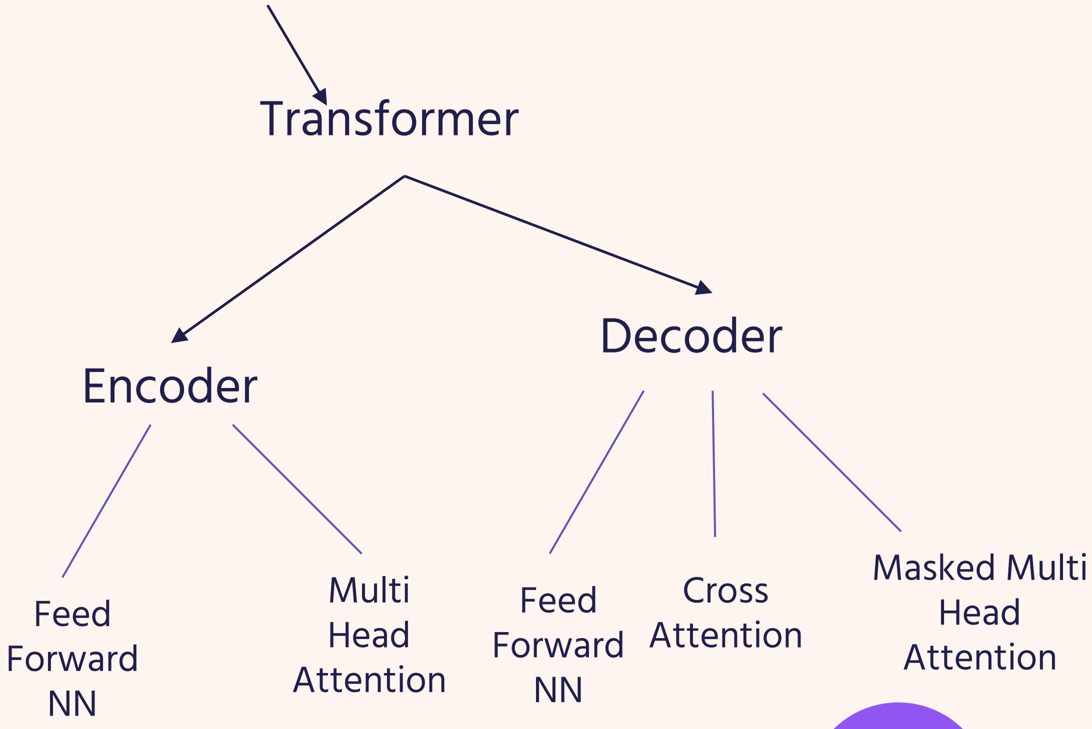
Search Engine Optimization

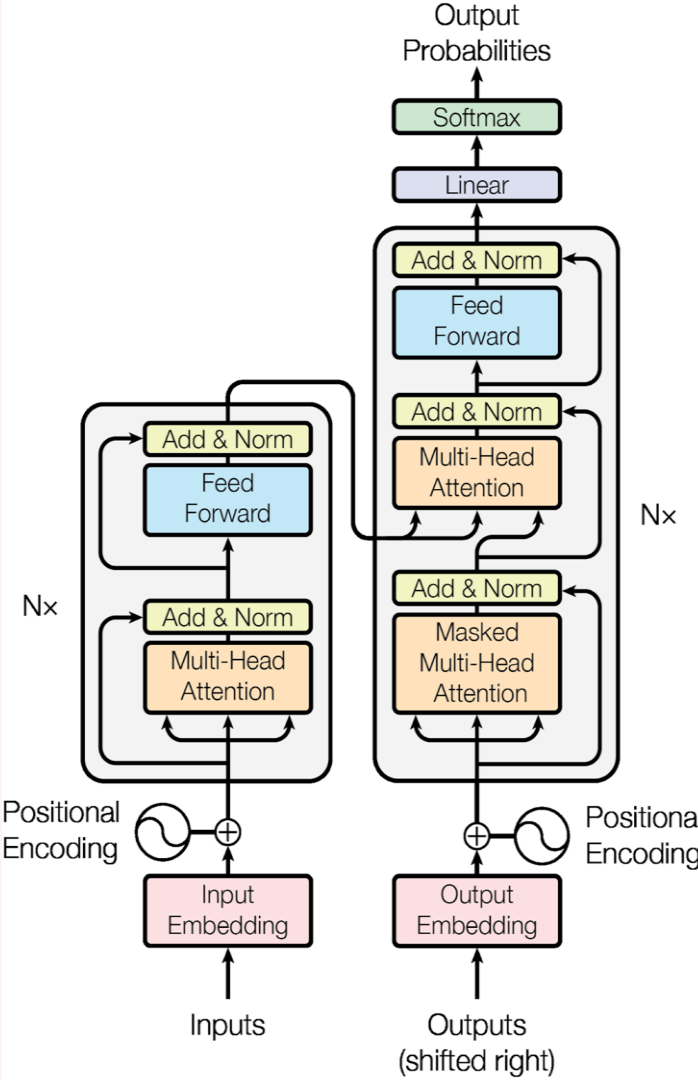
(e.g., Bing)

Challenges in XMC



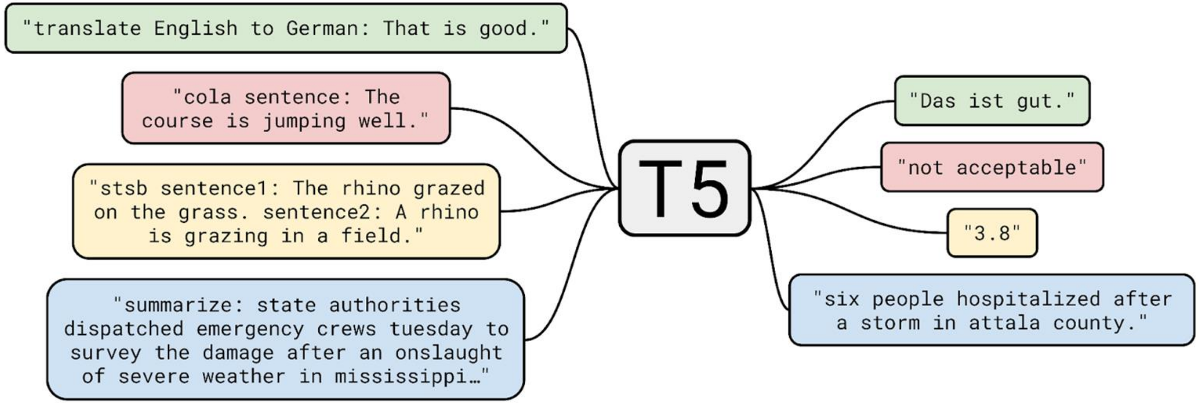
LLM


Transformer
Encoder
Decoder
Feed Forward NN
Multi Head Attention
Feed Forward NN
Cross Attention
Masked Multi Head Attention


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

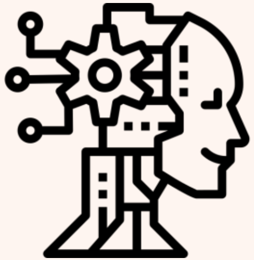
T5

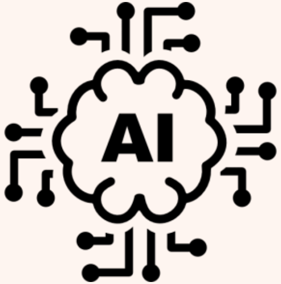
- Developed by Google
- T5 reimagines NLP as a simple text-to-text problem
- This approach allows to handle a variety of NLP tasks
- Unified Framework, Scalability, Performance, Flexibility



Solved a Real World Problem

Problem Statement: Attribute-Value Prediction From
E-Commerce Product Descriptions





Problem Statement

- Problem Overview: Predicting attribute-value pairs from unstructured product descriptions is a key challenge in e-commerce catalogs like Amazon, Walmart, and Alibaba.

- Task Objective: Develop a model that can automatically predict attribute-value pairs for a given product description.

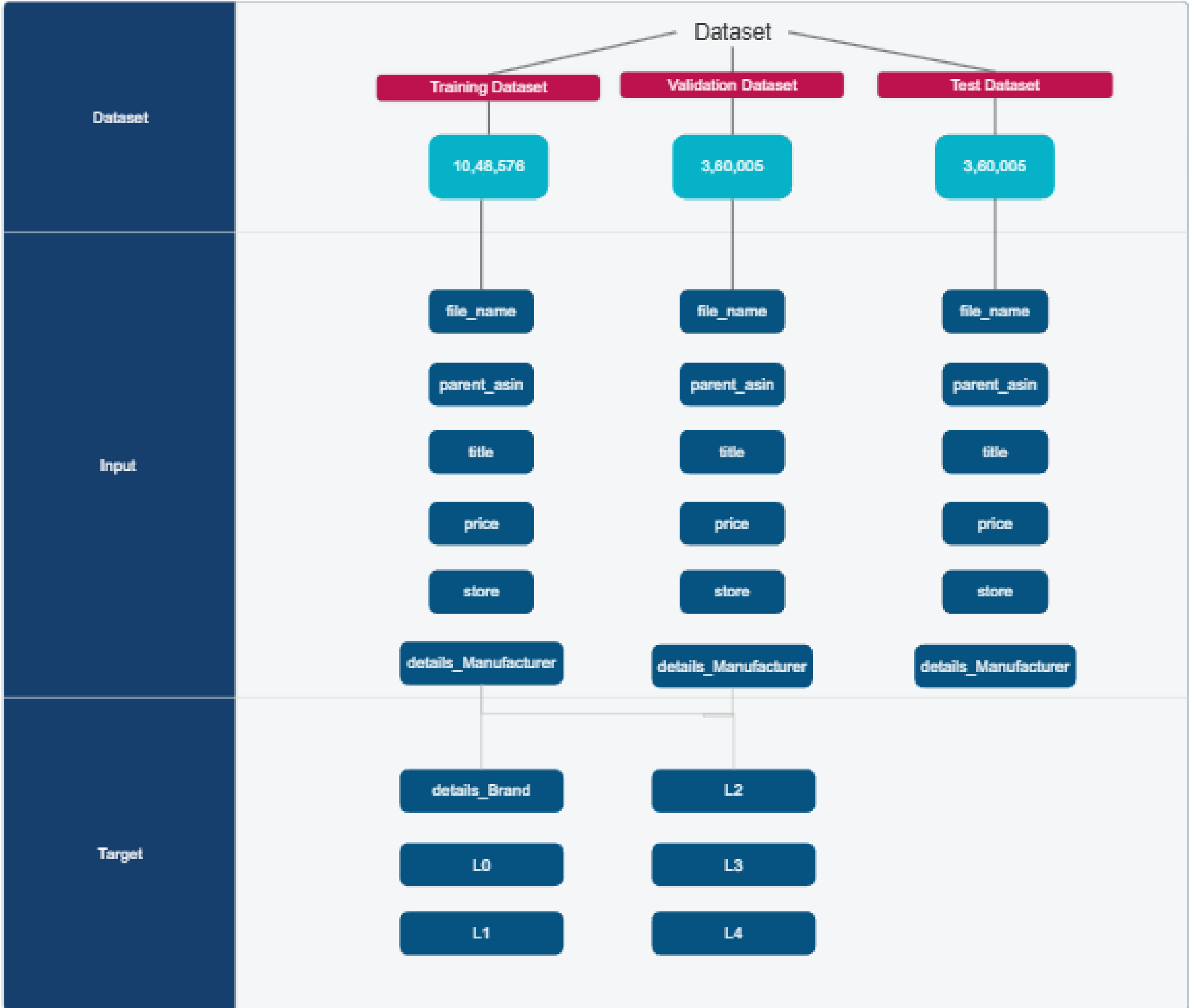
- Input Provided:
  - Short product title
  - Product description
  - Details about the store and manufacturer (optional for use)

Problem Statement

- Prediction Requirements:
  - Predict the values for 5 levels of product categories (from L0 to L4)
  - Predict the brand for the given product

- Prediction Goal:
  - Enable automatic categorization and brand identification from unstructured product information, enhancing product listing accuracy and search functionality.

Dataset Analysis



Input Features

S. No.	File_name	Product_ID	Title	Price	Store	Manufacturer
1	Automotive	B000C9FJ1Y	GM 15-8535 Heating and Air Conditioning Blower...	NaN	GM	ACDelco
2	Electronics	B075HRNB8K	Polaroid PIF-300 Instant Film - Twin Pack	NaN	Polaroid	Polaroid
3	Automotive	B07J9ZFLT8	CoolingCare Radiator for 1992-2004 Chevy GMC C...	NaN	Cooling Care	Cooling Care
4	Electronics	B086384SF5	RM-GD014 Remote Control Replacement for Sony ...	7.95	Elekpia	Elekpia Factory
...	...	...	...	...	...	...
1680013	Home_and_Kitchen	B004L8V95C	Urnex Cafiza Espresso Machine Cleaning Tablets...	7.34	Urnex	Urnex

Output Features

S. No.	Brand_category	L0_category	L1_category	L2_category	L3_category	L4_category
1	GM (1934)	Automotive (484633)	Replacement Parts (253381)	Engine Cooling & Climate Control (18199)	Heating (2081)	Blower Motors (1718)
2	Polaroid (213)	Electronics (166486)	Camera & Photo (25077)	Film Photography (1030)	Film (413)	na (858841)
3	Cooling Care (62)	Automotive (484633)	Replacement Parts (253381)	Engine Cooling & Climate Control(18199)	Radiators (4484)	na (858841)
...	...	...	...	...	...	...
1680012	Remo (143)	Instrument Accessories (48)	Drum & Percussion Accessories (48)	Concert Percussion Accessories (48)	Drum Accessories (48)	Snare Drum Accessories (22)
1680013	Urnex (32)	Kitchen & Dining (54)	Small Appliance Parts & Accessories (54)	Coffee & Espresso Machine Parts & Accessories(54)	Coffee & Espresso Machine Cleaning Products (32)	na (858841)
Total Labels	14262	31	198	911	2199	1852
Tail Labels	1752	0	3	47	133	133

Results

Metric	details_Brand	L0_category	L1_category	L2_category	L3_category	L4_category
Accuracy	0.995	0.85	0.739	0.805	0.692	0.59
Precision	0.967	0.886	0.758	0.777	0.71	0.544
Recall	0.98	0.791	0.778	0.719	0.722	0.515
F1 Score	0.987	0.824	0.787	0.743	0.684	0.587

The Tail Label Problem

These issues arise from the uneven distribution of labels or classes in the dataset, where a number of labels (frequent labels) dominate the majority of data points, leaving other labels (rare labels) sparsely represented. It leads to:

Bias Towards Head Labels- Models tend to focus on the "head labels," which are frequently occurring categories, at the expense of tail labels. This imbalance exacerbates the retrieval inefficiency for niche items, which can be critical in applications like product recommendation or content tagging.

Evaluation Metrics

Label	Per-Class F1 Score	Macro-Averaged F1 Score
Airplane	0.67	$\frac{0.67 + 0.40 + 0.67}{3}$ = **0.58**
Boat	0.40	
Car	0.67	

Evaluation Metrics

Label	True Positive (TP)	False Positive (FP)	False Negative (FN)	Micro-Averaged F1 Score
Airplane	2	1	1	$\frac{TP}{TP+\frac{1}{2}(FP+FN)} = \frac{6}{6+\frac{1}{2}(4+4)}$ = **0.60**
Boat	1	3	0	
Car	3	0	3	
TOTAL	**6**	**4**	**4**	

Label	Per-Class F1 Score	Support	Support Proportion	Weighted Average F1 Score
Airplane	0.67	3	0.3	$(0.67 * 0.3) + (0.40 * 0.1) + (0.67 * 0.6)$ = **0.64**
Boat	0.40	1	0.1	
Car	0.67	6	0.6	
Total	-	10	1.0	

Roadmap of future work

Literature review, dataset exploration, and baseline evaluations.

Framework implementation, customization, and experimentation.

Comprehensive evaluation, documentation, and real world applications

END

Thanks

Do you have any questions?

ics