Technischer Bericht: Datenvorverarbeitung

Titel: Vorverarbeitung und statistische Analyse des NABU-Prädatoren-Datensatzes

Datum: 01.Dezember 2025

Autor: SURK GOUN JANG

1 Einleitung und Zielsetzung

Ziel dieses Arbeitsschrittes war die Strukturierung und Bereinigung der von NABU bereitgestellten Rohdaten („Urdaten“). Die ursprünglichen Daten lagen in einer stark verschachtelten Ordnerstruktur vor und enthielten zahlreiche redundante Serienaufnahmen (Bursts), die das Training von Machine-Learning-Modellen negativ beeinflussen können (Gefahr des Overfitting).

2 Methodik

Zur Bereinigung des Datensatzes wurden zwei Python-gestützte Verfahren angewendet:

2.1 **Struktur-Flattening**: Rekrusive Extraktion aller Bilddateien aus Unterordnern in eine flache Klassenstruktur.

2.2 **Duplikat-Filtering (Deduplication)**:

2.2.1 **Zeit-Kriterium:** Bilder, die innerhalb eines Zeitfensters von 1 Sekunden aufgenommen wurden, wurden als zusammengehöriges Ereignis gruppiert.

2.2.2 **Qualitäts-Kriterium**: Mittels des Laplace-Varianz wurde aus jeder Gruppe nur das qualitativ beste Bild („Best Shot“) extrahiert.

3 Ergebnisse der Datenbereinigung

Insgesamt wurden **3.277 Objektes** aus den Rohdaten analysiert. Nach der Filterung verbleiben **2.503 Einzelbilder** für das Training. Dies entspricht einer Reduktion des Datenvolumens um **ca. 23,6%**.

Tabelle 1: Detaillierte Übersicht der Klassenverteilung

Rang	Klasse (Tierart)	Rohdaten (Anzahl)	Nach Selektion	Reduktion	Anmerkung
1	**Fuchs**	**1274**	**993**	**-22,1%**	**Dominante Klasse**
2	Krähen	322	295	-8,4%	
3	Kolkrabe	331	294	-11,2%	
4	Marderhund	261	214	-18,0%	
5	**Seeadler**	**366**	**133**	**-63,7%**	**Starke Reduktion**
6	Iltis	180	128	-28,9%	
7	Rind	79	70	-11,4%	
8	Steinmarder	57	53	-7,0%	
9	Silbermöwe	57	52	-8,8%	
10	Steinwälzer	50	47	-6,0%	
11	Lachmöwe	46	46	0,0%	Kein Redundanz
12	Rohrweihe	47	45	-4,3%	
13	Sturmmöwe	44	42	-4,5%	
14	Igel	41	33	-19,5%	
15	**Austernfischer**	**85**	**29**	**-65.9%**	**Massive Reduktion**
16	Dachs	25	25	0,0%	Kein Redundanz
17	Heringsmöwe	30	24	-20,0%	
18	Habicht	19	15	-21,1%	
19	Mäusebussard	13	12	-7,7%	
20	**Hund**	**3**	**3**	**0,0%**	**Zu wenig Daten**
21	**Hermelin**	**3**	**3**	**0,0%**	**Zu wenig Daten**
22	**Schmarotzerraubmöwe**	**3**	**3**	**0,0%**	**Zu wenig Daten**
23	**Lachseeschwalbe**	**2**	**2**	**0,0%**	**Zu wenig Daten**
	Gesamt	**3.277**	**2.503**	**-23.6%**	**Finaler Datensatz**

4 Analyse und Beobachtungen

4.1 **Klassen-Ungleichgewicht (Class Imbalance)**: Der Datensatz weist ein extremes Ungleichgewicht auf. Der Fuchs dominiert mit fast 1.000 Bildern (ca. 40% des Datensatzes), während 14 der 23 Klassen weniger als 50 Bilder besitzen. Dies stellt eine Herausforderung für das Modelltraining dar.

4.2 **Effektivität des Burst-Filters**: Bei Arten wie dem Seeadler (Reduktion -64%) und dem Austernfischer (Reduktion -66%) zeigte der Filter die größte Wirkung. Dies lässt darauf schließen, dass diese Tiere tendenziell länger vor der Kamera verweilen, was ohne Filterung zu einer massiven Datenredundanz geführt hätte.

4.3 **Kritische Klassen**: Arten mit weniger als 10 Bildern (Hermelin, Lachseeschwalbe, Hund, Schmarotzerraubmöwe) verfügen über zu wenig Varianz für ein stabiles Training und sollten gegebenenfalls für die Klassifikation gruppiert oder ausgeschlossen werden.

5 Fazit

Die Datenvorverarbeitung war erfolgreich. Der resultierende Datensatz ist nun frei von redundanten Serienaufnahmen und technisch bereinigt.