Dataset Comparison Report

Generated on: 2025-12-11 14:56:41

Executive Summary

This report compares two memory forensics datasets used for malware/spyware detection:

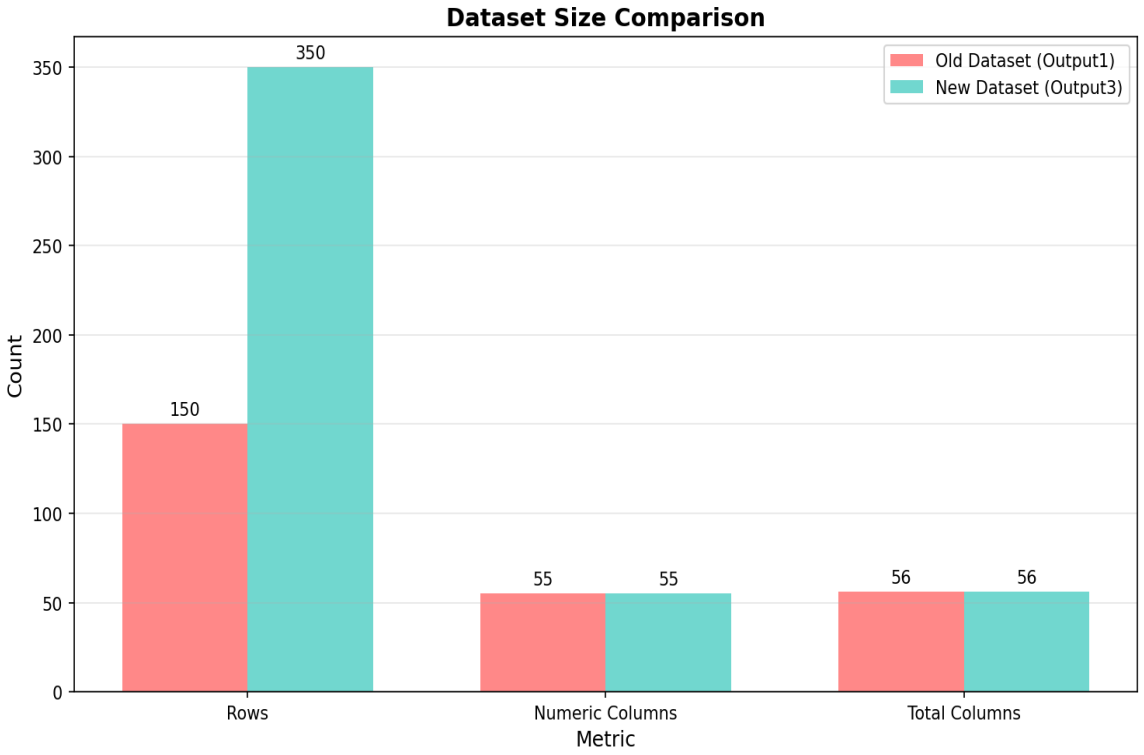
Old Dataset (Output1.csv): 150 samples with 56 features
New Dataset (Output3.csv): 350 samples with 56 features

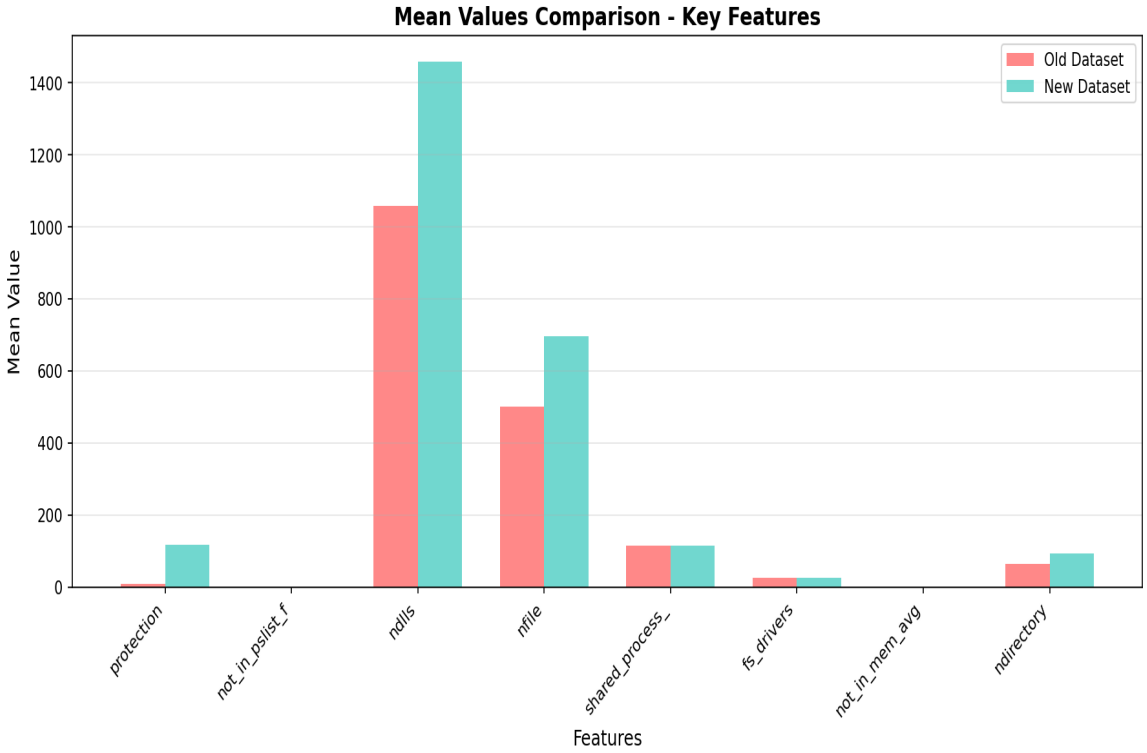
Key Findings:
• Sample count difference: 200 (133.3% increase)
• Both datasets contain 55 numeric features for analysis
• Feature structure is consistent between datasets

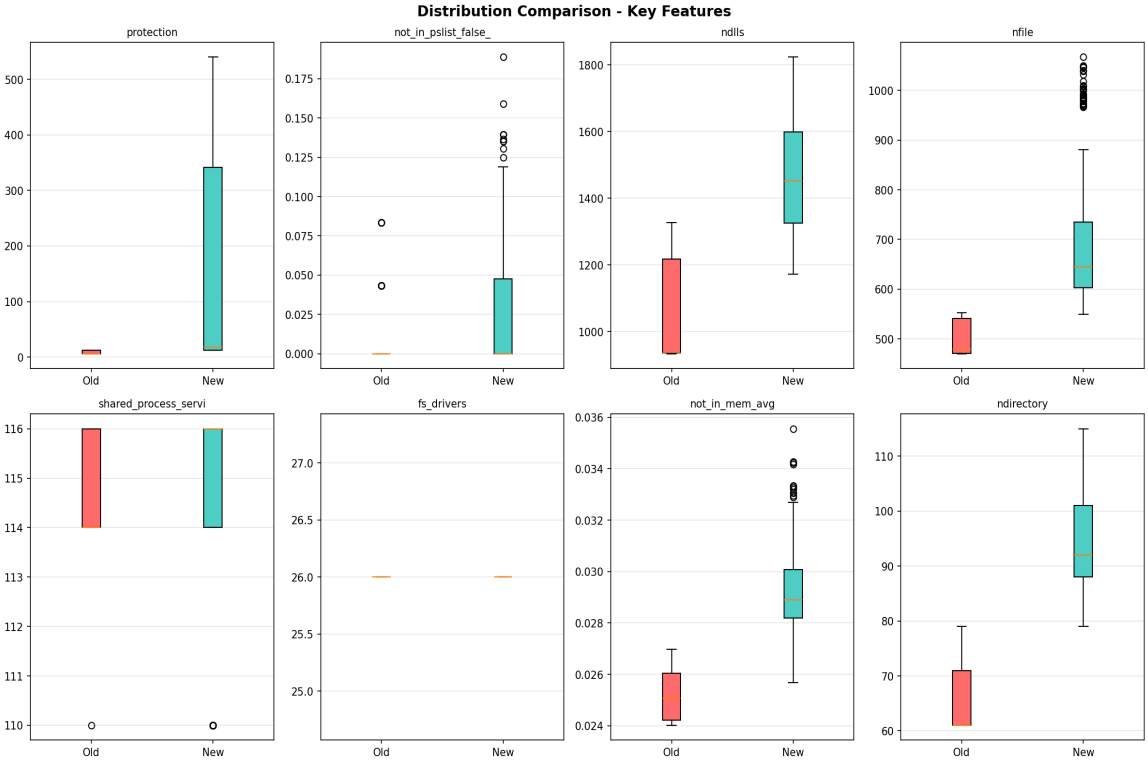
Dataset Overview

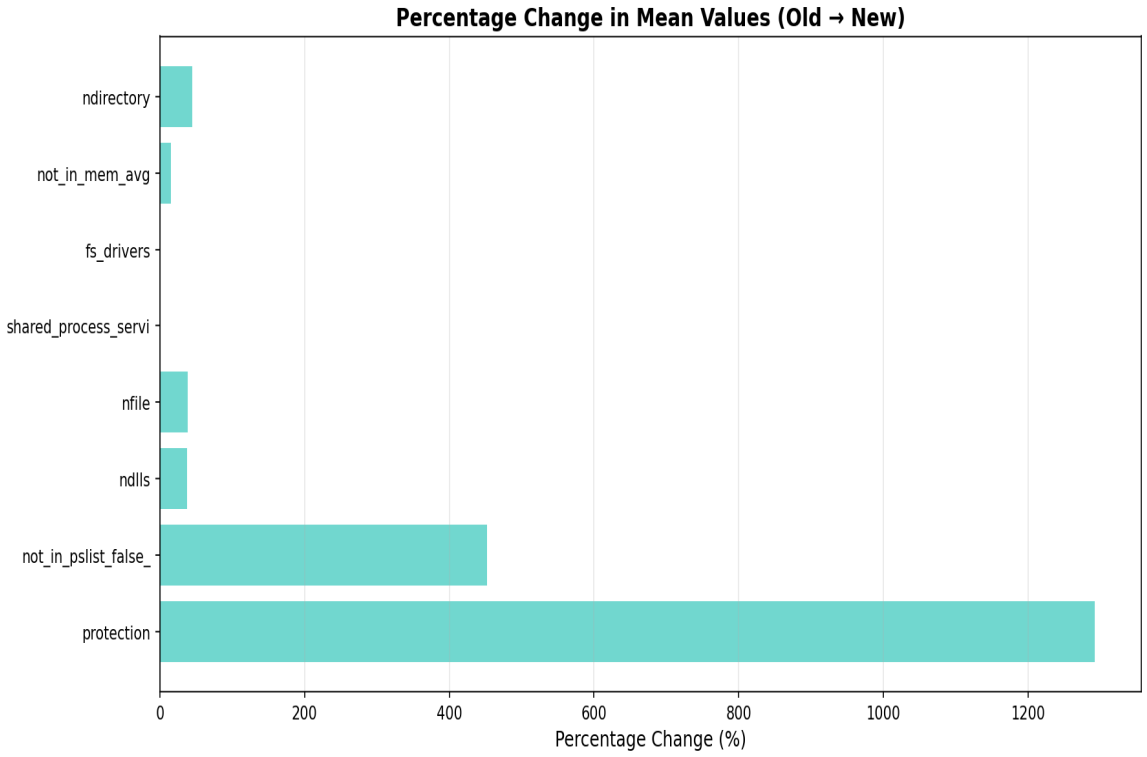
Metric	Old Dataset (Output1)	New Dataset (Output3)	Difference
Total Rows	150	350	200
Total Columns	56	56	0
Numeric Columns	55	55	0
Missing Values	0	0	0

Visual Comparison






Distribution Comparison - Key Features
protection
not_in_pslist_false_
ndlls
nfile
shared_process_servi
fs_drivers
not_in_mem_avg
ndirectory
Old
New


Percentage Change in Mean Values (Old → New)
ndirectory
not_in_mem_avg
fs_drivers
shared_process_servi
nfile
ndlls
not_in_pslist_false_
protection
0
200
400
600
800
1000
1200
Percentage Change (%)

Statistical Comparison - Key Features

Feature	Old Mean	New Mean	Old Std	New Std	Change %
protection	8.40	116.91	2.95	162.13	+1291.8%
not_in_pslist_fals	0.01	0.03	0.02	0.04	+452.3%
ndlls	1058.51	1457.51	149.51	170.10	+37.7%
nfile	501.89	695.26	33.48	138.62	+38.5%
shared_process_ser	114.79	115.06	1.06	1.60	+0.2%
fs_drivers	26.00	26.00	0.00	0.00	+0.0%
not_in_mem_avg	0.03	0.03	0.00	0.00	+15.5%
ndirectory	65.40	94.55	5.44	9.00	+44.6%
not_in_load	27.33	43.62	2.98	5.17	+59.6%
not_in_mem	27.33	43.62	2.98	5.17	+59.6%

Conclusions

Dataset Evolution Analysis:

1. **Sample Size:** The new dataset contains 200 more samples (133.3% increase), indicating expanded data collection.

2. **Feature Consistency:** Both datasets maintain the same feature structure with 56 columns, ensuring compatibility for model training and comparison.

3. **Data Quality:** Missing values are minimal in both datasets, with the new dataset having 0 missing values compared to 0 in the old dataset.

4. **Statistical Variations:** The statistical distributions show variations between datasets, which may reflect different malware samples, system states, or collection periods.

Recommendations:
• Consider combining both datasets for more robust model training
• Validate model performance on both old and new data separately
• Monitor feature drift over time for production systems