Efficiency and Accuracy Assessment of an Automated Image Classifier for Commercially Important Fish Species

May 2026

Authorship: This report was produced by NOAA's Chief Economist Team at NOAA's Performance, Risk and Social Science Office (PRSSO) in collaboration with NOAA's Southeast Fisheries Science Center (SEFSC). The collaboration included Arnab Acharya[a], Sara Guenther[b], Jeffrey Wielgus[a], Monica Grasso[a], George Gardner[a], Ravi Chittila[a], Joseph Conran[a], Tadesse Wodajo[a], Tony Wilhelm[a], Sara Thomas[c], Matthew Campbell[d], Kelsey Martin[d], and Jack Prior[d].

[a]PRSSO, Office of the Chief Finance Officer (OCFO), National Oceanic and Atmospheric Administration (NOAA), 1335 East-West Highway, Silver Spring, MD, 20910.
[b]NOAA Affiliate, PRSSO/OCFO and Montana State University, 732 W Grant Street, Bozeman, MT, 59717.
[c]NOAA Affiliate with Technical and Engineering Support, Alliance, SEFSC, National Marine Fisheries Service (NMFS), NOAA, 3500 Delwood Beach Road, Panama City, FL, 32408.
[d]SEFSC/NMFS, NOAA, 3209 Frederic Street Pascagoula, MS, 39567.

Send all correspondence to Arnab Acharya at arnab.acharya@noaa.gov.

Acknowledgements and Disclaimer: We would like to thank Brett Alger for providing written comments, and Shawn Klimek along with the participants of The Ebb & Flow Brown Bag Workshop at the Department of Commerce. We also thank Patricia Kelly for editorial assistance. All errors are the responsibilities of the authors.

ABSTRACT

Experts that analyze video collected from underwater optical systems can routinely provide accurate species counts. However, the high opportunity costs of expert time have entailed delays in processing large amounts of video data. Artificial Intelligence (AI) has the potential to reduce these delays and costs. To establish a baseline for this potential, we conducted a formative evaluation of a model used to count fish by species, run in the Visual and Image Analytics for Multiple Environments (VIAME) software. We compared the results from an application of the software against those generated by human experts across 357 Gulf of America underwater video deployments. The tool excelled at minimizing false-positive identifications. However, it consistently undercounted target species that were actually present. The missed identification was prevalent, particularly in cases where there was a high-degree of species diversity within the video. Automated processing time was substantially longer than the time spent for manual review. Fish density and species diversity increased processing time for both the manual reviewer and the automated classifier. In general, the statistical explanation for variations in processing time for manual review was more robust than the VIAME tool. We conclude that operating multiple continuously running machines and improving the VIAME software to become more accurate at distinguishing fish across species, especially in dense and speciose videos, presents a promising pathway to achieve automated survey efficiency.

Key Words: Visual and Image Analytics for Multiple Environments (VIAME) software, Evaluation, Comparing methods

1. Introduction

Evaluating the size and health of fish populations provides the foundational science required to set sustainable catch limits and prevent overfishing. As part of the support for the fishery industry, amounting to $319 billion in 2022, the National Oceanic and Atmospheric Administration (NOAA, 2026) has had continued fishery-independent surveys over three decades. NOAA and partners launched the Gulf Fishery Independent Survey of Habitat and Ecosystem Resources (G-FISHER) program in 2020 to monitor reef fishes through use of video deployment in conjunction with automated image recognition tools.

Under the G-FISHER program (Switzer *et al.*, 2023), NOAA's Southeast Fisheries Science Center (SEFSC) deploys underwater video cameras at roughly 1,000 representative monitoring stations in the Gulf each year. Manually processing this massive volume of imagery is costly and time-consuming and induces delays between data collection and the stock assessment with implications for data quality. Human experts must analyze hours of footage and meticulously identify and count individual species across a wide range of environmental and biological conditions. Fish density is highly variable, with some stations having only a few species present while others have dense, highly mobile schools.

Automated tools that could classify species would free up time for other activities and address delays in their availability for other projects. To address these challenges, NOAA leadership has prioritized the adoption of artificial intelligence (AI) and machine learning (ML) techniques (Richards *et al.* 2019). If accurate and either fast or relatively inexpensive to run, an AI/ML tool is generally being seen as extremely promising (Haque and Jufaili, 2026). However, such factors as long processing time and frequent needs to adjust the imaging software may also be costly.

To evaluate the potential of the image processing technology, SEFSC partnered with the Performance, Risk and Social Science Office (PRSSO) to conduct a pilot study of the Visual and Image Analytics for Multiple Environments (NOAA, VIAME) technology, which was developed by Kitware, Inc., in cooperation with NOAA. This study serves as a formative evaluation—an early-stage, evidence-building effort designed to assess the baseline feasibility and operational characteristics of this tool prior to full-scale implementation, including running VIAME across multiple computers running continuously.

Specifically, we investigated both the expediency (processing time) and reliability (accuracy) of the tool relative to expert manual reviews, while also examining how environmental conditions such as water clarity and habitat complexity affect its outputs. The analysis examined only the output from an application of the tool and did not address the underlying specific technical aspects of the tool that might be affecting the results.[1]

The evaluation was structured around several specific guiding questions:

- Where does the automated classifier perform as well as, or better than human analysts?

[1] For an assessment of the tool from a data analytics point of view, see Prior et al. (2023). These authors state that "automated image analysis through computer vision algorithms has emerged as a tool for fisheries to address big data needs, reduce human intervention, lower costs, and improve timeliness. Processing can occur 24/7 while employing human expertise may be intermittent and lengthen the time of completion."

- Where does the automated classifier perform poorly compared with human analysts?

By answering these questions, this report seeks to provide actionable insights for developing effective human-AI collaboration strategies in fisheries data processing.

2. Materials and Methods

2.1 Data Collection and Manual Review

The paper presents an evaluation of a tool to assess whether it is feasible, appropriate, and acceptable before it is fully implemented. This type of evaluation is called formative evaluation (CDC 2024, OMB, M-20-12, 2020).[2] The data used here is what experts and the automated produced as final outputs. The evaluation results do not address any specific design features that drive the VIAME model to recognize fish images. It is assumed standard tools were employed to test the validity of the automated tools by the software writers using the training data. The paper evaluates the quality of products from an AI application against the quality of the exact outputs produced by humans; such comparisons are new (see Cowley et al 2022 and Rainio et al. 2024).[3]

Data used in this study, were sourced from the 2024 Gulf Fishery Independent Survey of Habitat and Ecosystem Resources (G-FISHER, 2025). The data was collected using cameras that were deployed at monitoring stations across the Gulf of America, capturing 20–30 minutes of footage per site. For this analysis, 357 deployments from the year 2024 were examined, split between the eastern (301 videos) and western Gulf (56 videos), divided by the 89°W longitude line.[4] Deployments generally spanned the months of April to August. The survey uses a randomly stratified sample from 21 designated habitat strata; there are no hypotheses regarding how these strata affect different fish counts or processing times (Switzer *et al*. 2023).

There are five lenses that capture recording per camera. Each lens is evaluated for habitat conditions and readability, and one of the five lenses is selected for analysis. The lens with the best representative habitat for the station is the lens that is selected for processing.

For the current data, human experts with at least seven years of experience in the relevant field conducted the manual reviews, identifying total fish counts, species richness, and individual species counts. Individual species were counted for their MaxN, or mincount, which is the maximum number of individuals observed in a frame at any given time during the 20-minute read period. This manual process includes a second expert to verify 10% of the first counts; the current verification process reported an agreement rate of 97%. Both sets of experts follow a G-FISHER standard operating procedure (SOP) to count and identify the fish and the second readers were informed that the video had already been examined. Experts were advised to record the number of minutes it took them to complete a video read, which was specifically done for

[2] Accessed February 2025.

[3] Authors note that examining different measurement tools has a long history, especially in epidemiology. Some supervised machine learning tools, such as Receiving Operating Characteristic curves, can be applied.

[4] The higher sites from the eastern part of the Gulf was the result of timing of the completion of video review.

this analysis and is not part of the G-FISHER SOP. These manual counts serve as the baseline for this assessment. Experts also ranked two environmental variables: Relative Visibility (1–9 scale)[5] and Habitat Complexity (2–10 scale).[6]

2.2 Automated System (VIAME)

The automated system, currently at version 2.5, utilizes a convolutional neural network (CNN) architecture trained on 737,430 images of 180 fish species. This study focuses on 10 commercially important species. The tool's sensitivity is governed by a confidence threshold (0.1 to 0.95, incremented by 0.1 until the level 0.9), which is designed to minimize Type I errors (false positives)—defined here as identifying a species when it is not present or misidentifying a non-fish object as one of the ten species.

The model is run on the entire length of video and predictions are generated for each fish track at the frame level. This level of resolution supports the extraction of the necessary information to directly compare counts as though a manual reader were to have read the video. In other words, the same MaxN method of counting is able to be extracted once the exported files are fed into a separate R script (Rv.4.5.0). Additional information including how long the predictions took to run on each video is also obtained.

2.3 Evaluation Framework

The evaluation utilized 1,900 species-deployment observations (after excluding 19 outliers) from 357 deployments. We assessed two primary metrics, expediency and reliability. Analyses used descriptive and inferential statistics. Expediency was evaluated by comparing manual processing times (MPT) against the automated processing times (APT) that were spent to record the information from the 357 videos. Reliability was assessed by comparing automated counts against the manual baseline.

3. Results

3.1 Processing Time and Efficiency

Human experts record a manual processing time while the automated tool records a similar time duration for how long it took to process all the frames from the video. A significant disparity was observed between MPT and APT. As summarized in Table 1, the mean MPT was 36 minutes (range: 7–132), while the mean APT was considerably higher at 167 minutes (range: 68–1143). The durations would matter if results need to be made available quickly and tools are costly to run. Therefore, processing times from both methods need to be examined.

Distributions of these times (Z-scores standardized for mean and standard deviation for the samples for each of the tools), presented in Figures 1a-b, showed that APT values were clustered between the 25th to 75th percentile with high value outliers. MPT and APT were weakly

[5] Visibility varies due to suspended matter in the water column, light levels, or other sources of variation in the depth-of-field. Higher numbers indicate worse visibility.

[6] Habitat complexity is generated by biotic and abiotic features. Complexity increases with higher numbers. The value is the sum of structural and biotic complexity, each indexed 1-5.

correlated (r = 0.391), a value that dropped to 0.211 when 15% of the data were trimmed from both sides. Figure 2 shows no particular directional relation between APT and MPT. The lack of similar trends by deployment between the two processing times is hard to interpret; however, it may suggest human expertise is employed in different ways for recognizing species than the automated tool.

One would expect a similar time-spent gradient for the same type of video environment across the two methods. The gradient of time spent across the different video environments for the two processing times should be highly correlated; however, there is little correlation across the strata as rank-correlation across the strata for the two types of processing times was 0.442. Figure 3 shows the 21 strata do not affect times in any particular way, although the two strata AML and AMS, which are artificial strata show up as consuming more time.[7] The strata analysis offers mixed conclusions regarding the similarities between the times. The correlation should be higher as the two tools measure the exact same phenomenon. It is also possible that the strata do not affect processing times at all, a possibility somewhat confirmed for manual count later in the document.

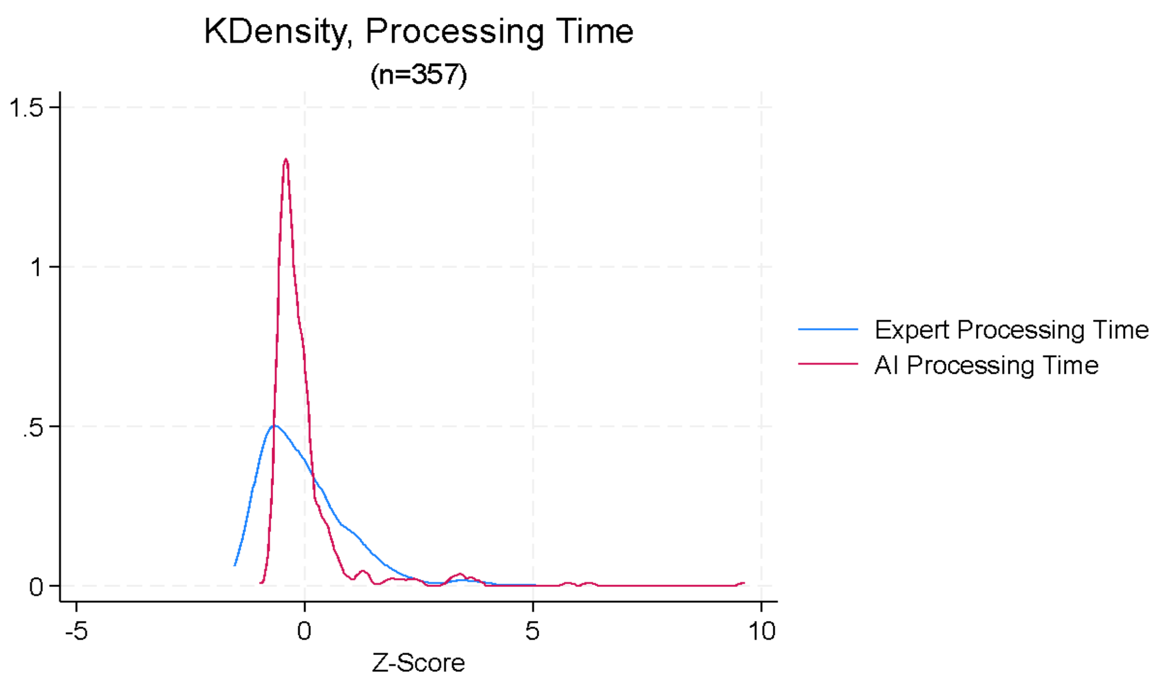

Figure 1a: KDensity, Processing Time

[7] Artificial habitats have denser abundance of fish (Karnauskas *et al.*, 2017).

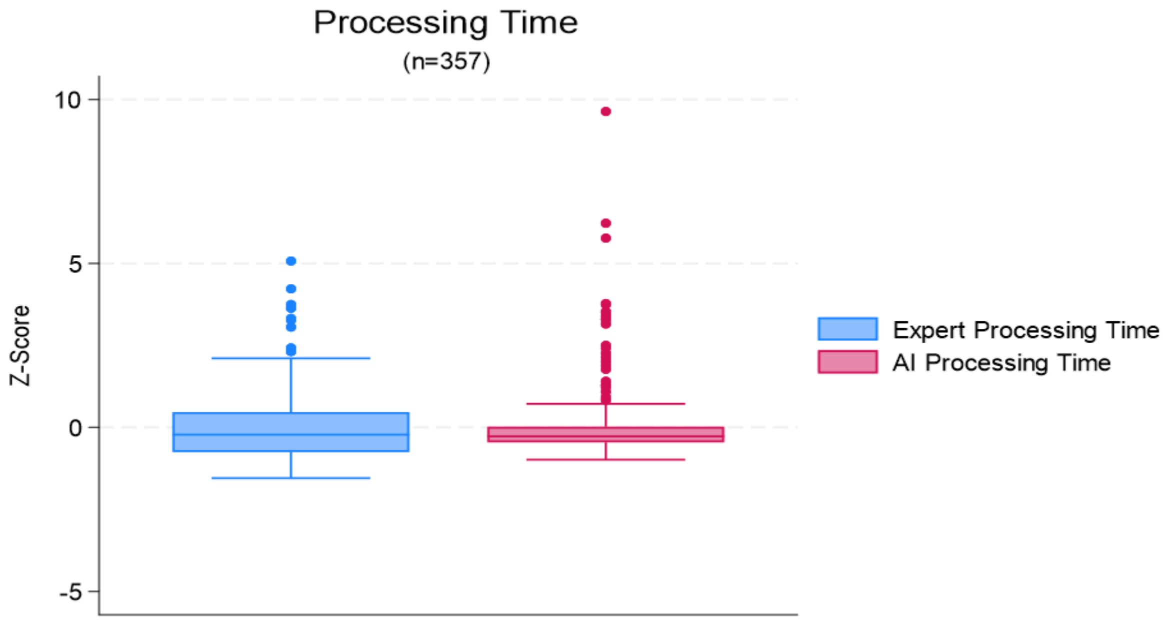

Figure 1b: Processing Time

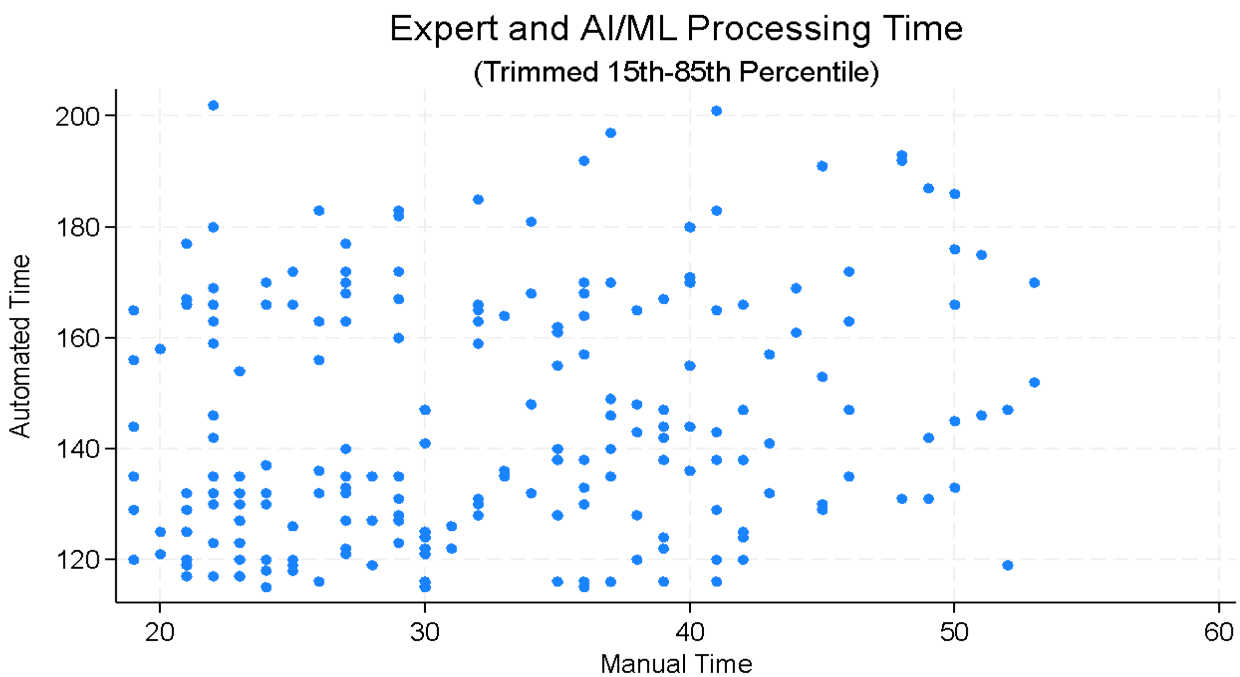

Figure 2: Expert and AI/ML Processing Time

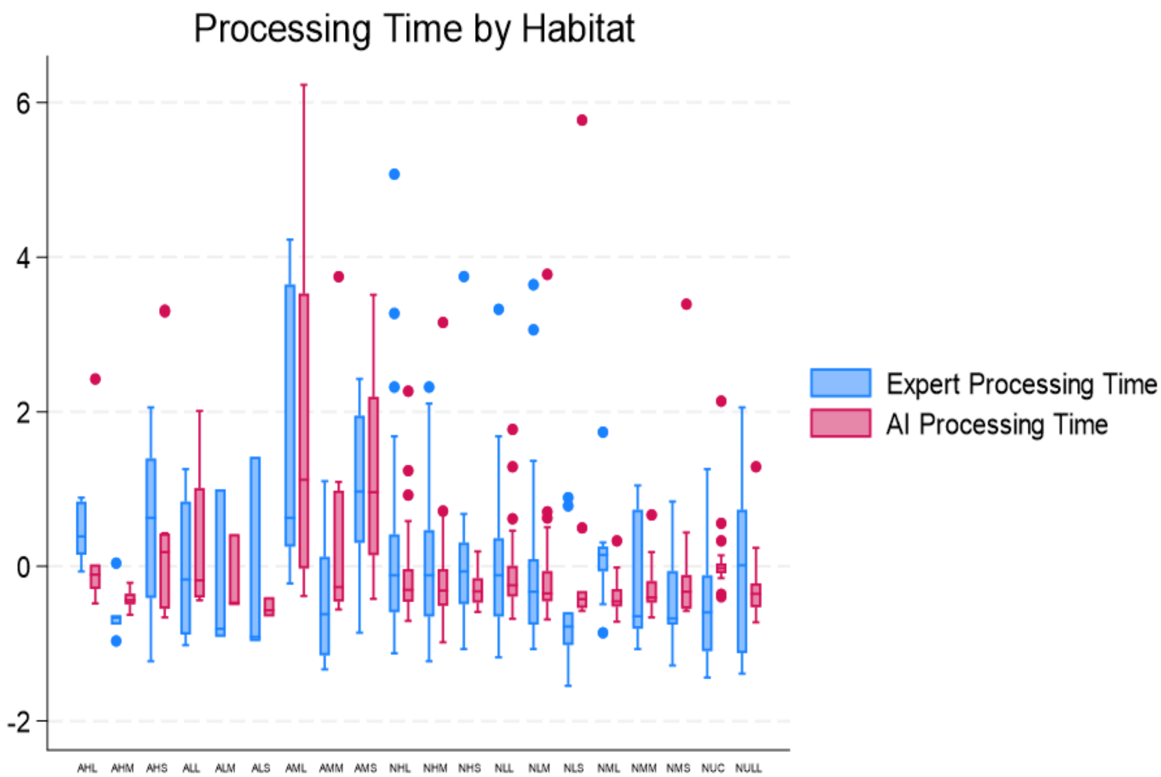

Figure 3: Processing Time by Habitat

Table 1: Processing Time and Deployment Description (n=256)

Variables	Mean	Standard Deviation	Minimum	Maximum
Manual Processing Time	36.21	18.88	7	132
Automated Processing Time	167.58	101.23	68	1,143
Total Fish	107.9	296.08	0	2,076
Total Species	8.46	4.88	0	23
Relative Visibility (1-9)	5.75	1.15	3	9
Habitat Complexity (2-7)	3.44	1.1	2	7

Summary values for factors across the deployments used in explaining the variation in the processing times. Manual and automated processing times differ with automated times having larger means, range and variation.

Table 2: Correlation: Features of the Deployment and Processing Times

Variable	Total Fish	Total Species	Relative Visibility	Habitat Comp	Manual Time	Automated Time
Total Fish	1					
Total Species	0.192	1				
Relative Visibility	-0.018	-0.304	1			
Habitat Complexity	0.039	0.516	-0.230	1		
Manual Processing Time	0.307	0.608	-0.055	0.363	1	
Automated Processing Time	0.616	0.156	0.017	0.015	0.391	1

Reports the correlation between factors used to explain the variations in the processing times.

3.2 Factors Influencing Processing Time

Regression analyses (Table 3) revealed that environmental factors present in the video are able to explain the variations in MPT, while for APT, explanatory power is much lower. There is greater understanding as to the factors affecting MPT than we have for APT. The expected impact of these factors are the following conditioned upon all factors other than the specific factors remaining constant:

- Relative Visibility is a ranking of increasing difficulty in visibility; thus, it would increase processing time.
- Habitat Complexity is a ranking of increasing complexity of the physical environment present in the video. It would increase the processing time.
- Total Fish is the total count of the fish made by expert readers. As the numbers of total fish count increase, the videos would become more complex. This factor would increase processing time.
- Total Species, defined as the number of unique species identified by human experts in the video, will influence processing time. A higher count of species increases complexity and, consequently, increases processing time.[8]
- It is possible that the strata introduced as a fixed effect can affect the results greatly.

The following is the summary as to how the factors affect processing time where any significance level at or below 0.1 was observed:

[8] The paper also hypothesizes that these factors also influence accuracy.

- Videos from the eastern part of the Gulf see a lower processing time for APT and MPT. This factor is significant for all estimations.
- As visibility becomes poorer, processing time increases; statistical significance was only observed for MPT.
- Habitat complexity increased processing time for MPT, but not for APT. Results were similar for both the dummy and demeaned fixed-effect manual time. The non fixed-effect model explained MPT more strongly, shown through a higher R-Square. The complexity variable became significant when the visibility variable was dropped for APT; however, this was in the opposite direction than expected. There is likely to be only a weak relationship between habitat complexity on any of the processing time, as when the strata averages are taken out, the variable becomes slightly less significant for MPT.
- The number of total fish and different species have expected impacts in all cases, with statistical significance.

The demeaned fixed effect model adds insight into how the strata might be affecting processing time (column 3 and 7). As would be expected the strata fixed effect shows the intercept term at a higher value for MPT. With unexplained strata factors removed, the results show that within each stratum variations are captured by the regression intercept. It is unusual to see the constant coefficient reduced as shown for APT particularly when it has the expected effect on MPT. This suggests that automated processing time is not specified properly although the same structure is used for MPT.[9] The analyses provide few results as to which factor to focus on in order to reduce APT. This would be of particular concern, from a cost perspective, if the use of the tool is energy-intensive when used in multiple computers. However, a start could be to see how the tool could be made faster for higher density of fish and diversity of species.[10] Although the strata effect is present, the results do not change much when a fixed effect for strata is employed (see estimations 1 and 3, Table 3).

[9] As dummy variable fixed effect models produce artificially higher R-square, the demeaned fixed-effect value shows the explanatory power (R-square) is much lower for the APT than it is for the MPT.
[10] A technical issue is that the residuals for the automated tools regression cannot at all be considered normal, raising concerns as to how predictable AI results are using standard statistical analyses.

Table 3: Determining Factors to Explain Variations in Processing Times

Variables	Manual Processing Time			Automated Processing Time				
	(1)	**(2)**	**(3)**	**(4)**	**(6)**	**(7)**	**(8)**	**(9)**
Region, East	-4.043* (2.089)	8.994*** (2.138)	-4.246** (1.961)	-32.511*** (10.374)	-53.056*** (14.247)	-25.049** (9,882)	-32.445*** (10.199)	-33.553*** (9.925)
Relative Visibility	2.113*** (0.647)	1.927*** (0.523)	2.055*** (0.694)	1.976 (2.599)	0.522 (3.336)	1.274 (2.867)	2.230 (2.608)	
Habitat Complexity	1.657** (0.747)	1.638* (0.855)	1.766* (0.775)	-3.381 (4.451)	-1.865 (4.241)	-1.113 (3.752)		-3.580*** (4.456)
Total Fish	0.013*** (0.004)	0.008*** (0.0026)	0.008** (0.003)	0.208*** (0.053)	0.178*** (0.049)	0.178*** (0.048)	0.209*** (0.053)	0.209 (0.053)
Total Species	2.273*** (0.202)	2.442*** (0.251)	2.334*** (0.194)	2.234* (1.328)	1.751* (1.750)	2.158 (1.399)	1.850* (1.043)	2.141*** (1.320)
Intercept	1.156 (4.681)	11.666** (5.406)	3.580*** (1.872)	**153.874***** (23.638)	191.728*** (31.626)	**21.12**** (8,910)	143.896*** (21.205)	167.551 (12.890)
R-square	0.436		0.414	0.395		0.327	0.395	0.396
Number of observations	357	357	357	357	357	357	357	357

*** p<.01, ** p<.05, * p<.1 Parenthesis: Standard Error Fixed Effect: New Habitat Strata from G-FISHER program. Columns 2 and 6 are fixed effect models with strata dummies; columns 3 and 7 report strata demeaned fixed effects. Determinants for automated times are few.

3.3 Accuracy and Confidence Thresholds

Accuracy is defined as the agreement between these two counting methods:

- An exact match occurs when the automated count agrees with the manual count for a species-deployment pair.
- An error (or an inaccurate match) is any automated count value that differs from the corresponding manual value.

The tool's accuracy is highly dependent on the confidence threshold (defined above). Table 4 describes the data by species. The manual count shows the presence of a higher number of total fish than the automated counts; for the species counts this is not always true. Focusing on the higher half of the threshold level, at a 0.6 threshold, the tool has only slight tendencies to overcount fish; as the threshold increased, this shifted toward significant undercounting (Figure 4). Figure 4 reports on differences in counts for species-deployment pairs.

The highest exact match rate (62%) occurred at thresholds of 0.7 and 0.8. When the error tolerance was expanded to ±1 fish, agreement rose to 85% at thresholds 0.6 and 0.7 (Figure 5a). More detailed analyses were conducted for the threshold level of 0.7, as this shows the highest level of accuracy.

At the 0.7 threshold, the automated tool showed a pervasive tendency to undercount species. Using a Bland-Altman plot, a standard tool to compare two measurements of the same phenomenon, (Figure 6) confirmed this, showing that differences (Automated - Manual) were almost exclusively negative or result in undercounting by the automated tools (Bland and Altman, 1986; Ryan and Woodall, 2005). The curve shows the differences between the two methods at the average values from both methods.

The number of deployments with none of the selected species present totaled 965 from the manual count, more than 50%. The corresponding number of deployments with one of the selected species present was 935. Accuracy also varied based on whether fish were actually present:

- Zero-presence of a species (Figure 5b): When manual counts were zero (965 observations) for a particular species, the tool was highly accurate, reaching a 99% exact match at high thresholds.
- Positive Presence of a species (Figure 5c): When a fish species was present (935 observations), exact matches dropped significantly. At the 0.7 threshold, the tool identified roughly half the total fish recorded by humans (Table 4). There was an exact match of 33%; and for a near match (matched number of ≤ 1), the corresponding value was 71%. Interestingly, as the threshold level is increased both exact and near matches decrease for the non-zero species count. This may be an indication that the tool is generating higher false negatives for the alternative hypothesis that there is a fish of the given species as the threshold level is increased.

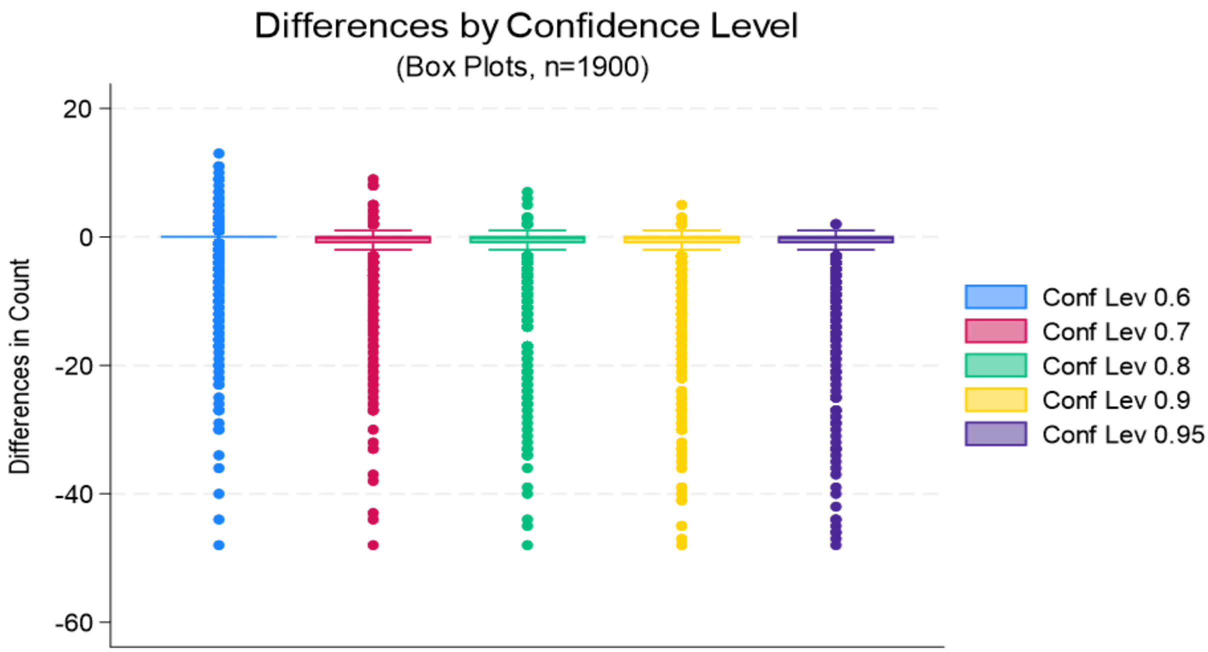

Figure 4: Difference: Automated Count - Manual Count, by deployment species pair

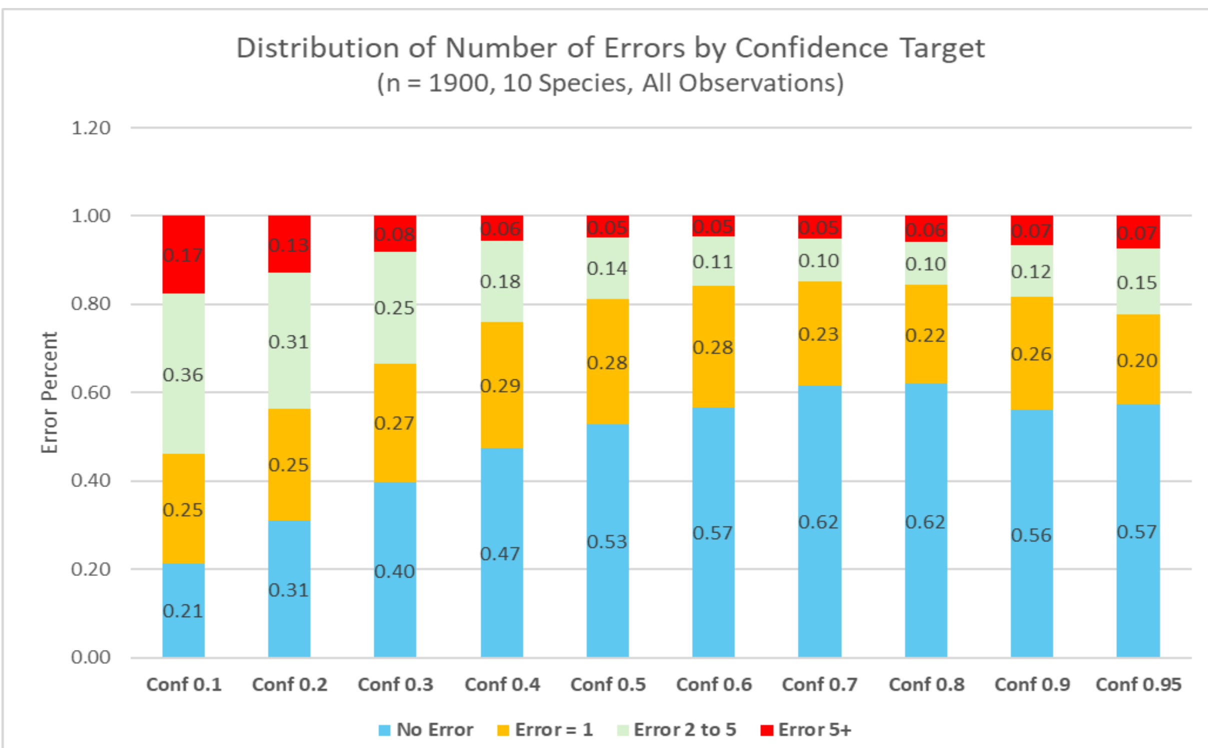

Figure 5a: Error Distribution, All Observations and Thresholds

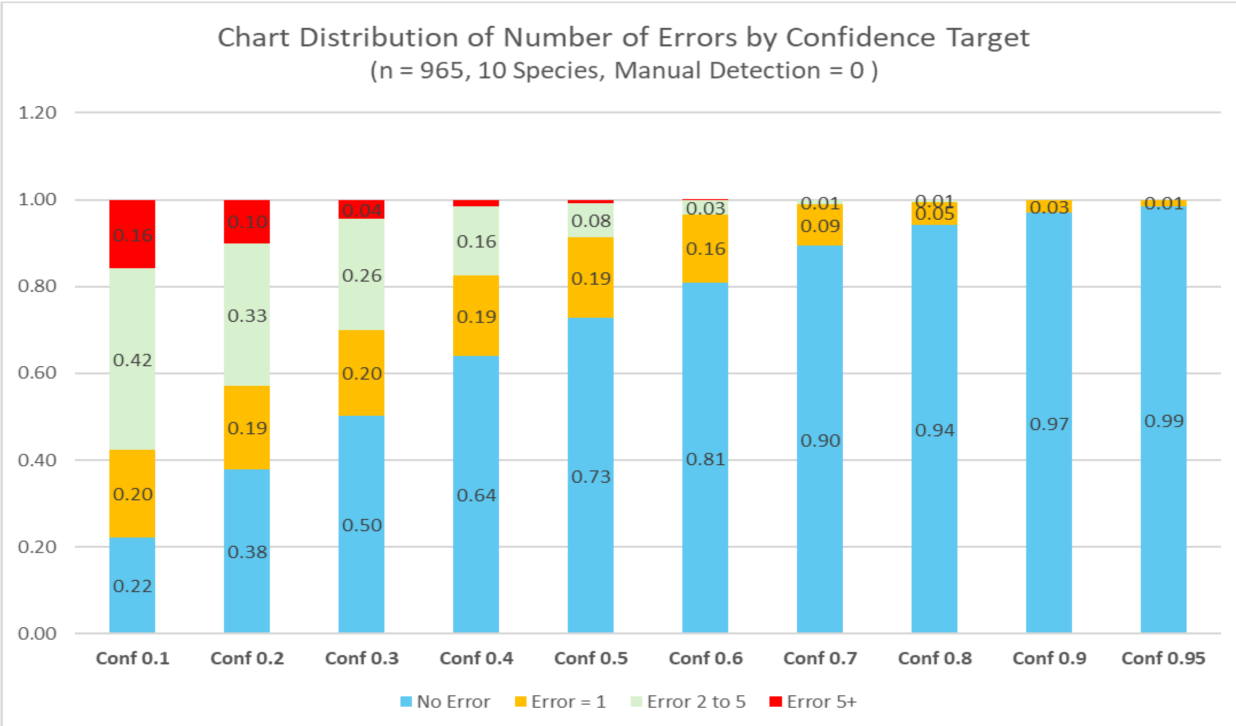

Figure 5b: Error Distribution, Manual Count = 0, All Thresholds

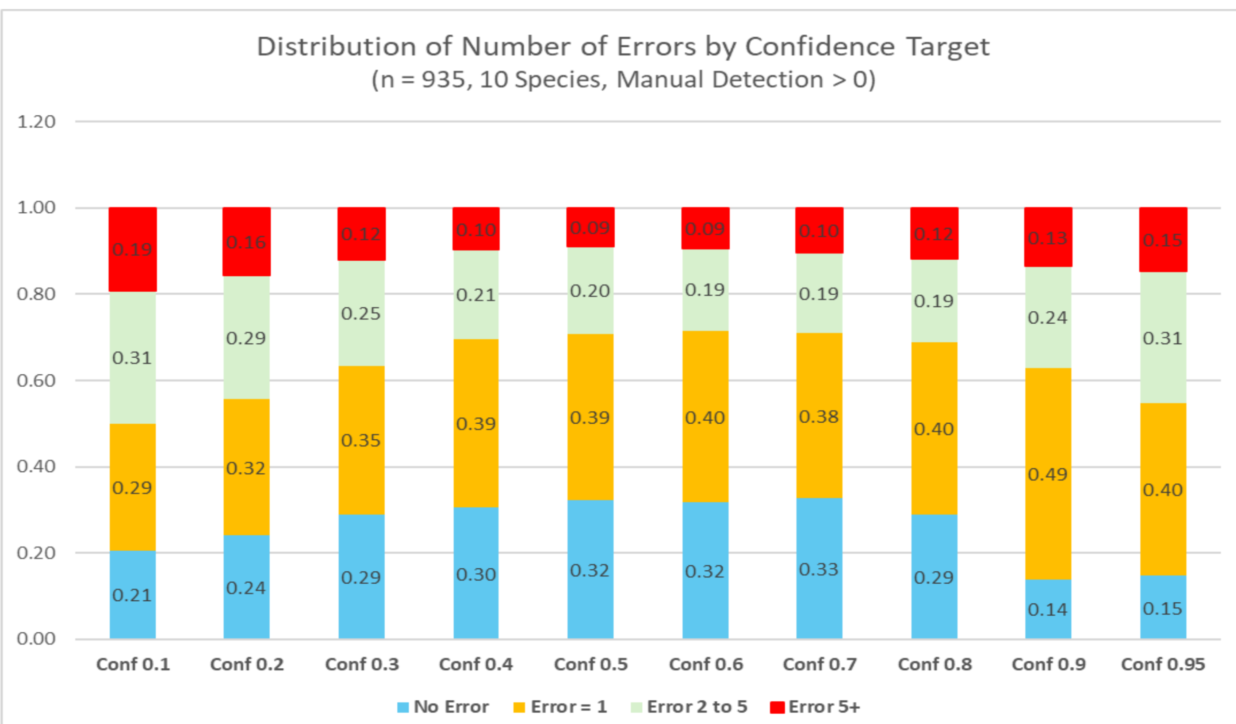

Figure 5c: Error Distribution, Manual Count > 0, All Thresholds

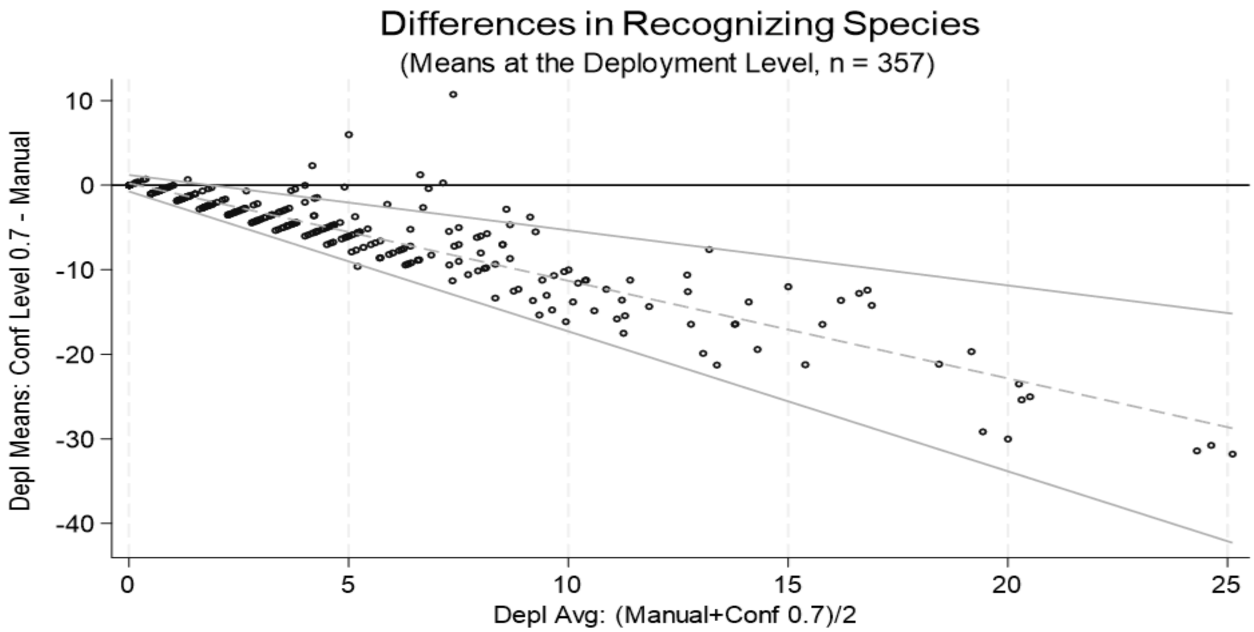

Figure 6: Automated Count - Manual Count, Bald-Altman Diagram

3.4 Species-Level Performance

Not all deployments had every species (Table 4). At the species level, accuracy was highest for gag grouper, *Mycteroperca microlepis* (80%), and lowest for lane snapper, *Lutjanus synagris* (44%,) (Figure 7a). Across all species, the tool was consistently more accurate in confirming the absence of a fish than recognizing that a species is present when there was actually a fish present (Figures 7b-c). Figure 7b shows that when a no count is seen at the manual level for a species, the automated tool reports a very high level of exact match at 91% on average across the species. However, the level of accuracy falls to 27% on average across the species when the manual count is greater than zero (Figure 7c).

Table 4: Aggregated Description of Species Count, 1900 species-deployment pairs

Type of Count	Number of Deployments Reporting Species			Total Number Counted by Species (In Reported Deployments)			
	Total Reported	Manual Count = 0	Manual Count > 0	Manual	Confidence Levels 0.6, 0.7 and 0.8		
Fish Species (Abbrev)	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Balistes capriscus (Bal cap)	329	196	133	276	344	307	271
Lutjanus campechanus (Lut cam)	354	151	203	947	1097	905	711
Lutjanus griseus (Lut gris)	151	82	69	377	153	113	77
Lutjanus synagris (Lut syn)	79	23	56	103	35	24	16
Mycteroperca microlepsis (Myc mic)	129	104	25	45	22	12	7
Mycteroperca phenax (Myc phe)	165	83	82	176	138	119	84
Pagrus pagrus (Pag pag)	216	85	131	491	399	356	316
Rhomboplites aurorubens (Rho aur)	250	135	115	1526	722	538	384
Seriola dumerili (Ser dum)	147	77	70	258	84	47	22
Seriola rivoliana (Ser riv)	80	29	51	104	63	51	35
Total Observations	1900	965	935	4303	3057	2472	1923

Columns numbered 1-3 indicate the number of deployment-species combinations for which a count is reported. Column 1 reports that for 329 deployment a Bal cap number is reported including where no fish is found. For 196 of the deployments. no Bal cap was found while for 133 deployment at least one Bal cap fish was found (Columns 2 and 3). Column 4 reports the manual count from all deployments; for Bal cap 276 total fish. Columns 5 to 7 report the number of fish observed by three different threshold levels.

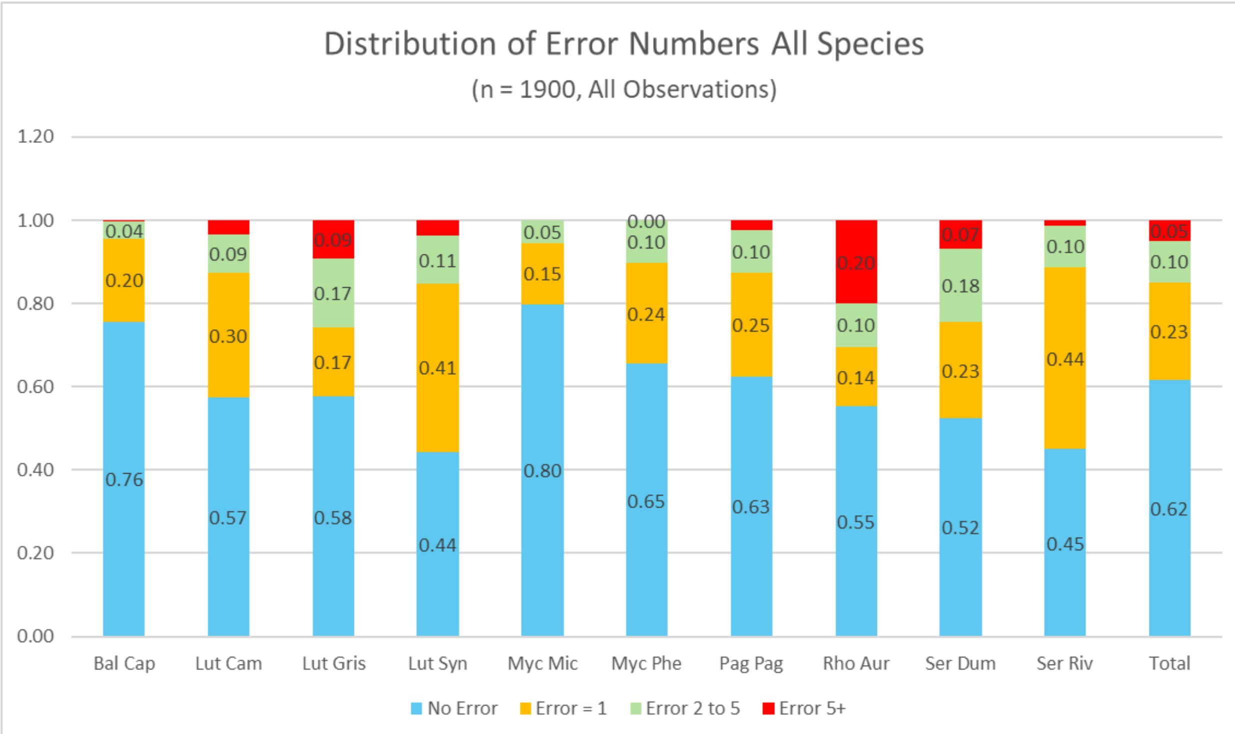

Figure 7a: Error Distribution by Species[a], Threshold Level 0.7

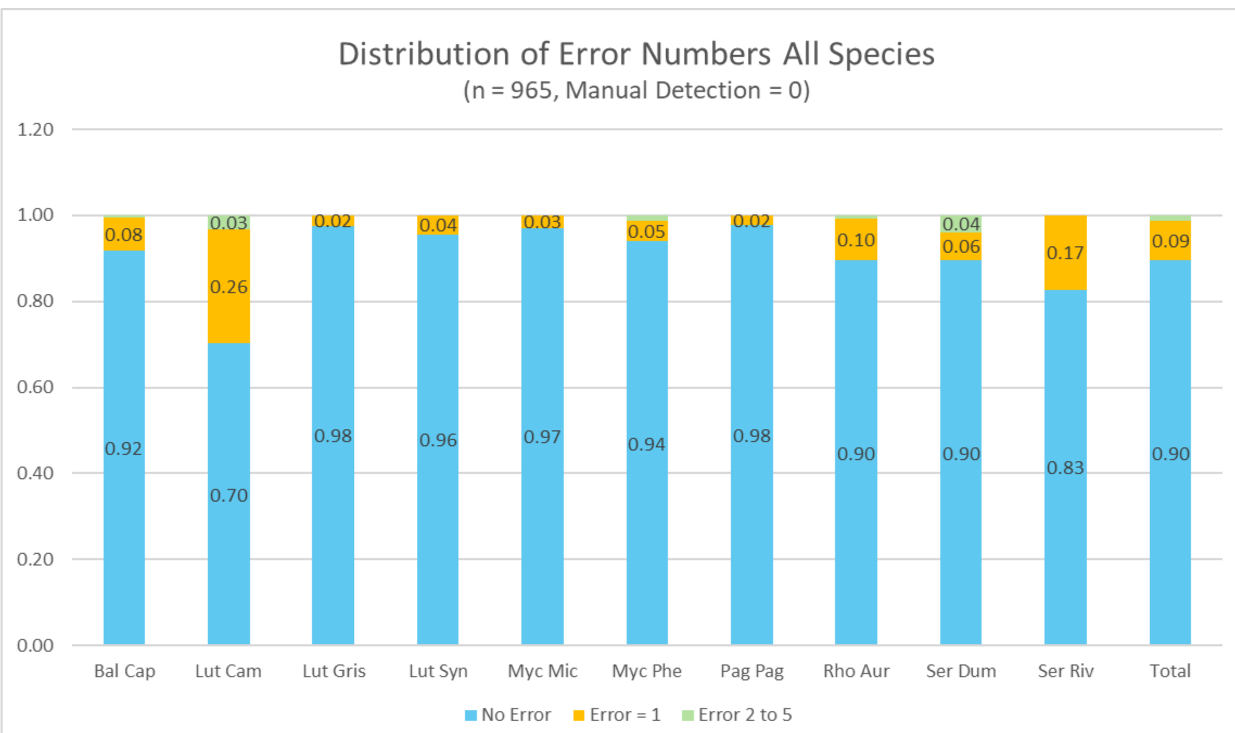

Figure 7b: Error Distribution by Species, Manual Count = 0, Threshold Level 0.7

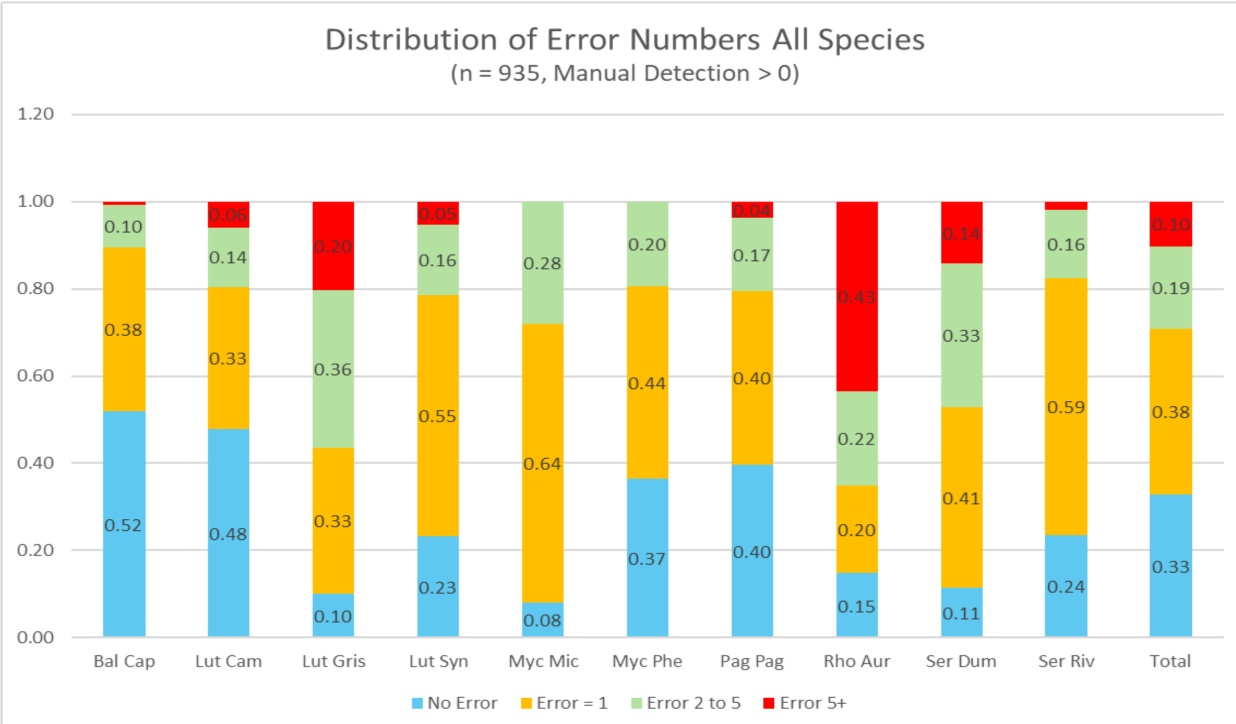

Figure 7c: Error Distribution by Species, Manual Count > 0

3.5. Factors Explaining Matching

Determining which video environmental features can impact matching and could improve the VIAME model. Logit regressions were run to find whether these features influence the overall rate of matching and by species. The following hypotheses, mostly similar to the factors used to explain processing times, were tested:

- Poor visibility and greater habitat complexity decreases the probability of an exact match.
- A greater number of fish in the deployment decreases the probability of exact match.
- Greater variation in the number of fish species decreases the probability of exact match.
- Artificial habitat strata (AML and AMS) could engender more errors (see previous result section).
- The Gulf regions (east and west) could affect the count.

Regression results are reported in Table 5. For the combined all species data (columns 1-2), a greater number of factors are significant, although the Pseudo R-Square value measuring explanatory power offers mixed results (columns 1-2). As the exact match value for cases where no fish is identified is quite high (see Figure 7b) including this factor as a dummy variable increases the explanatory power to 28% from a value less than 2%. As would be suspected from previous analyses, having no fish identified as one of the ten species under consideration reduces errors.

For the individual species, some regressions had a low number of observations. It is not clear whether this played a role in yielding results that lacked explanatory power. Explanatory power from these factors are very low as measured by pseudo R-square, ranging from 0.026 to 0.158 for regression by species and the coefficients for the factors may not be significantly different from 0 in most instances (Table 5, columns 3-12). The East region lowers exact matching for a few of the species and significant for two species and goes in opposite directions in those two instances (columns 3 and 12). Poor visibility generally lowers exact matching. Habitat complexity, when it is significant, reduces exact matching. Having more fish reduces exact matching in most cases. Having more species always reduces exact matching. The following observations can be made:

- Region: The East region generally reduces exact matching for a few of the species and is significant for only two species, though with opposite effects – lowers and increases matching probability for *Balistes capriscus* and *Seriola rivoliana*, respectively.
- Visibility: Poor visibility, when statistically significant, lowers exact matching.
- Habitat Complexity: When significant, habitat complexity reduces exact matching.
- Fish Count: Larger fish counts have no impact on exact matching.
- Species Counts: Having more distinct species always reduces exact matching.
- Artificial Strata: Artificial strata (2 strata) lowers accuracy.

Regression results (not reported) are nearly the same when the exact matching is relaxed to near matching (error ≤ 1). The current finding of low-level explanatory power does not allow the use of such tools as Receiving Operating Characteristics or Confusion Matrix to examine the predictive power of environmental factors present in the video (Cowley, 2022).

Table 5: Dependent Variable: Exact Match by Species from a Deployment/Species Observation[a]

	Combined Species[a]		Bal Cap	Lut Cam	Lut Gris	Lut Syn
Variables	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Region, East	-0.07 (0.161)	-0.25 (0.194)	-2.25** (1.037)	-0.24 (0.334)	-0.21 (0.191)	-0.09 (0.892)
Relative Visibility	-0.15*** (0.047)	-0.11* (0.057)	0.15 (0.145)	-0.26** (0.103)	-0.05 (0.042)	-0.42 (0.269)
Habitat Complexity	-0.02 (0.052)	0.00 (0.063)	0.04 (0.152)	-0.04 (0.117)	0.05 (0.041)	0.19 (0.306)
Total Fish	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.001)
No. of Species	-0.09*** (0.012)	-0.03** (0.015)	-0.09** (0.034)	-0.04 (0.028)	-0.04*** (0.010)	-0.16** (0.071)
Strata, Artificial	-0.51** (0.212)	-0.55** (0.263)	-1.53** (0.598)	-0.67 (0.573)	-0.13 (0.144)	0.06 (0.907)
Manual Value = 0		2.81*** (0.129)				
Intercept	2.25*** (0.383)	0.48 (0.467)	3.30** (1.423)	2.47*** (0.831)	1.35*** (0.356)	3.29 (2.306)
Pseudo R2	0.032	0.283	0.121	0.021	0.120	0.082
Number of observations	1900	1900	329	354	151	79

*** p<.01, ** p<.05, * p<.1, Standard Errors within parenthesis

[a]Species name in full found in Table 4. [b]Observations for all species are combined.

The results show the likely factors within the videos that could explain the errors are not able to provide much explanatory information. We find very low values for pseudo R-square. Prominently, for the total errors/lack of exact match the only factor that matters is the lack of presence of the specific species. When the manual value shows no fish of the specific species the explanatory power increases significantly as seen through pseudo R-square. One is able to conclude through all 12 estimation results that the variations in achieving exact matches are not explained through factors that present in the video.

Table 5 (Cont): Dependent Exact Match by Species from a Deployment/Species Observation[a]

	Myc Mic	Mic Phe	Pag Pag	Rho Aur	Ser Dum	Ser Riv
Variables	(7)	(8)	(9)	(10)	(11)	(12)
Region, East	0.81 (0.769)	-0.20 (0.523)	-0.36 (0.506)	0.21 (0.431)	0.83 (0.577)	1.84* (1.012)
Relative Visibility	-0.21 (0.283)	-0.30* (0.151)	0.00 (0.143)	-0.28** (0.127)	0.06 (0.176)	-0.23 (0.284)
Habitat Complexity	0.05 (0.238)	-0.35** (0.177)	-0.16 (0.159)	0.00 (0.139)	0.16 (0.193)	-0.43 (0.284)
Total Fish	0.00 (0.001)	0.00 (0.001)	0.00 (0.001)	0.00 (0.000)	0.00 (0.000)	0.003** (0.001)
No. of Species	-0.15** (0.064)	-0.05 (0.041)	-0.04 (0.036)	-0.13*** (0.035)	-0.10** (0.043)	-0.03 (0.058)
Strata, Artificial	-0.63 (0.955)	0.32 (0.938)	0.61 (0.713)	-1.35* (0.702)	-0.18 (0.660)	-1.57 (1.110)
Intercept	3.16 (2.069)	4.35*** (1.335)	1.67 (1.145)	2.87*** (1.049)	-0.38 (1.413)	0.99 (2.131)
Pseudo R2	0.057	0.051	0.032	0.066	0.050	0.158
Number of observations	129	165	216	250	147	80

*** p<.01, ** p<.05, * p<.1, Standard Errors within parenthesis

[a]Species name in full found in Table 4.

One is able to conclude through all 12 estimation results that the variations in achieving exact matches are not explained through factors that present in the video.

4. Summary and Discussion

The objective of this formative evaluation was to determine if automated image recognition could reliably replicate the manual fish-counting baseline. This transition is essential; while human experts are highly accurate, the manual review process creates a significant data bottleneck and can delay fishery stock assessments, a foundational activity for fishery management. The VIAME is highly accurate in case when there are no species to identify; it has the potential to be a time saving technology. And the longer automated processing times in comparison to manual processing times could still already be cheaper. As the accuracy in the cases where there is a species to identify is poor, improvements to the technology need to take place. Perhaps the improvements would also be vastly time consuming. To improve the technology it is important to know what factors affect time and accuracy.

The current analyses show factors that would affect human processing times and accuracy do not affect automated processing times or accuracy. This is a concern as patterns that are observed for experts do not offer ready solutions for improving the VIAME. These results are summarized in Table 6. A notable result is that greater diversity of species decreases accuracy.

Table 6: Results Summary

Observed Influencing Factors	Manual Processing Time	Automated Processing Time	Accuracy for Different Species[a]
Regional Indicator	Shorter time for Eastern Region	Shorter time for Eastern Region	No effect in most cases
Relative Visibility	Poorer visibility increases time	No Effect	No effect or decreases accuracy
Habitat Complexity	Greater complexity increases time	No effect	No effect in most cases
Total Fish Present	More fish present increases time	More fish present increases time	No effect in most cases
Total Species	More types species increases time	Weak relation in increasing time	Decreases accuracy in most cases
Overall explanatory Factor (R2)	Acceptable Level	Acceptable Level	Unacceptable in most cases
Artificial Strata			Decreases accuracy or no impact

[a]The results for accuracy were reported by species and for all fishes.

The factors present in video are able to explain the variations that we find for the processing times through manual means. The variations observed for the outcomes of the automated processing method largely cannot be explained.

Specialists in the field must calibrate the tolerable inaccuracies although inaccuracies are likely to decrease with newer iterations of the tool. With tolerable accuracy levels, the tool holds promise if processing time is inexpensive. In the present case, with a Manual Processing Time of 36 minutes per deployment, expert time saved for the 357 deployments would be 214 hours, adding up to several thousands of US dollars in wages.

Our results indicated that using raw output from the current AI/ML models cannot be recommended at this point, and therefore significant human-in-the-loop effort is required to produce trustworthy results for use in natural resource management for current model iterations. Underlying aspects involved in model training were not considered here but are clearly important to consider moving forward. For instance, class imbalance is a known issue with the VIAME models utilized, and several different efforts have been made to deal with this issue in VIAME and should also be tested using this framework. Additionally, other model training efforts will soon be undertaken, such as hierarchical frameworks.

While results indicated that processing time was not enhanced using the automated tool, this result has to be considered relative to two important considerations. First, the automated tool generates frame level counts and classifications, whereas the human annotations integrate singular counts over the entire video annotation process. Thus, automated data sets are orders of magnitude richer in underlying detail. Accuracy notwithstanding, the clear advantage is that the automated datasets can be used to estimate a variety of metrics beyond a singular MaxN value, which human annotations could not produce without significantly higher effort to generate frame level data.

Second, the automated processing procedures were conducted on instances of VIAME that could only access single graphics processing units (GPU), located on desktops or in limited versions deployed in non GPU-enabled web environments. Current efforts in an associated Optics Strategic Initiative run by NOAA Fisheries will deploy VIAME in GPU-enabled cloud environments that will optimize workflows by accessing as many GPUs as possible under budget constraints. Thus, processing power resolves model efficiency; and the underlying model precision is of higher importance and consideration for future efforts. Imprecise models will require significant human-in-the-loop intervention despite clear advances in processing power and efficiency offered by cloud-based solutions.

The results demonstrate a fundamental difference in how humans and AI perceive environmental noise. Human processing time was significantly hindered by poor visibility and habitat complexity. Human readers could focus on these factors to improve their abilities. In contrast, the automated processing time remained largely unaffected by these environmental factors present in the video. A similar conclusion was made regarding the ability of AI tools to identify species of fish. Human achievements are more predictable than what we observe with the automated tool. Expanding the scope of evaluation to include other optical surveys within the SEFSC and other regional fisheries offices in future studies would provide a broader understanding of AI's applicability and effectiveness in fish stock assessments across diverse contexts.

References

Bland, J.M. and Altman, D.G. (1986) 'Statistical methods for assessing agreement between two methods of clinical measurement', *Lancet*, 1, pp. 307–310.

The Center for Disease Control (CDC) (2024) *CDC Approach to Program Evaluation*. Available at: https://www.cdc.gov/evaluation/php/about/index.html (Accessed: February 2026).

Cowley, H.P., Natter, M., Gray-Roncal, K. et al. (2022) 'A framework for rigorous evaluation of human performance in human and machine learning comparison studies', *Scientific Reports*, 12, p. 5444. doi: 10.1038/s41598-022-08078-3.

Haque, S.A. and Al Jufaili, S.M. (2026) 'Applications of Artificial Intelligence in Fisheries: From Data to Decisions', *Big Data and Cognitive Computing*, 10(19). doi: 10.3390/bdcc10010019.

Karnauskas, M., Walter, J.F., Campbell, M.D., Pollack, A.G., Drymon, J.M. and Powers, S. (2017) 'Red Snapper Distribution on Natural Habitats and Artificial Structures in the Northern Gulf of Mexico', *Marine and Coastal Fisheries*, 9(1), pp. 50–67. doi: 10.1080/19425120.2016.1255684.

The National Oceanic and Atmospheric Administration (NOAA) (2025) *Gulf Fishery Independent Survey of Habitat and Ecosystem Resources (G-FISHER), SEAMAP Spherical Camera Deployment and Video Reading Standardized Operating Protocol*. G-FISHER.

The National Oceanic and Atmospheric Administration (NOAA) (2026) *Fisheries Economics of the United States, 2023*. NOAA Technical Memorandum NMFS-F/SPO-254. National Marine Fisheries Service.

The National Oceanic and Atmospheric Administration (NOAA) (2026) *Artificial Intelligence and Advanced Sampling Instruments (AIASI) Software Platform*. Available at: https://www.st.nmfs.noaa.gov/aiasi/Software.html (Accessed: February 2026).

The Office of Management and Budget (OMB) (2020) *M-20-12: Phase 4 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Program Evaluation Standards and Practices*. Available at: https://www.whitehouse.gov/wp-content/uploads/2020/03/M-20-12.pdf (Accessed: March 2026).

Prior, J.H., Campbell, M.D., Dawkins, M., Mickle, P.F., Moorhead, R.J., Alaba, S.Y., Shah, C., Salisbury, J.R., Rademacher, K.R., Felts, A.P. and Wallace, F. (2023) 'Estimating precision and accuracy of automated video post-processing: A step towards implementation of AI/ML for optics-based fish sampling', *Frontiers in Marine Science*, 10, p. 1150651. doi: 10.3389/fmars.2023.1150651.

Rainio, O., Teuho, J. and Klén, R. (2024) 'Evaluation metrics and statistical tests for machine learning', *Scientific Reports*, 14, p. 6086. doi: 10.1038/s41598-024-56706-x.

Richards, B.L., Beijbom, O., Campbell, M.D., Clarke, M.E., Cutter, G., Dawkins, M., Edgington, D., Hart, D.R., Hill, M.C., Hoogs, A., Kriegman, D., Moreland, E.E., Oliver, T.A., Michaels, W.L., Piacentino, M., Rollo, A.K., Thompson, C., Wallace, F., Williams, I.D. and Williams, K. (2019) *Automated Analysis of Underwater Imagery: Accomplishments, Products, and Vision*. NOAA Technical Memorandum NOAA-TM-NMFS-PIFSC-83, 59 p. doi: 10.25923/0cwf-4714.

Ryan, T.P. and Woodall, W.H. (2005) 'The most-cited statistical papers', *Journal of Applied Statistics*, 32, pp. 461–474.

Switzer, T.S., Keenan, S.R., Thompson, K.A., Shea, C.P., Knapp, A.R., Campbell, M.D., Noble, B., Gardner, C. and Christman, M.C. (2023) 'Integrating assemblage structure and habitat mapping data into the design of a multispecies reef fish survey', *Marine and Coastal Fisheries*, 15(4), p. e210245. doi: 10.1002/mcf2.10245.