Human Factors and Simulation, Vol. 83, 2023, 254–264
https://doi.org/10.54941/ahfe1003946

Assessing the Impact of Automated Document Classification Decisions on Human Decision-Making

Mallory C. Stites, Breannan C. Howell, and Phillip Baxley

Sandia National Laboratories, Albuquerque, NM, 87185, USA

ABSTRACT

As machine learning (ML) algorithms are incorporated into more high-consequence domains, it is important to understand their impact on human decision-making. This need becomes particularly apparent when the goal is to augment performance rather than replace a human analyst. The derivative classification (DC) document review process is an area that is ripe for the application of such ML algorithms. In this process, derivative classifiers (DCs), who are technical experts in specialized topic areas, make decisions about a document's classification level and category by comparing the document with a classification guide. As the volume of documents to be reviewed continues to increase, and text analytics and other types of models become more accessible, it may be possible to incorporate automated classification suggestions to increase DC efficiency and accuracy. However, care must be taken to ensure that model-generated suggestions do not introduce unacceptable errors into the process, which could lead to disastrous impacts for national security. In the current study, we assessed the impact of model-generated classification suggestions on DC accuracy, response time, and confidence while reviewing document snippets in a controlled environment and compared them to DC performance in the absence of a model (baseline). Across two assessments, we found that correct model suggestions improved human accuracy relative to baseline, and increased speed of response relative to baseline when full-length documents were used. Incorrect model suggestions produced a higher human error rate (for short but not full-length documents), especially when model explanations were provided. Incorrect suggestions also elicited longer responses for unclassified documents. DCs reported higher confidence when they complied with incorrect suggestions from an interactive model, relative to cases in which they correctly disagreed with them. These results highlight that although ML models can enhance performance when the output is accurate, they may impair analyst decision-making performance if inaccurate. This has the potential for negative impacts on national security. Findings have implications for the incorporation of ML or other automated suggestions not only in the derivative classification domain, but also in other high-consequence domains. The effects of model accuracy and amount of information displayed from the model should be taken into account when designing automated decision aids.

Keywords: Machine learning (ML) interaction, Human decision-making, Document classification

INTRODUCTION

Derivative classifiers (DCs) perform the difficult task of identifying classification sensitivities in documents and ensuring that they are marked appropriately to prevent the unintentional release of classified information. The volume of information needing DC review continues to increase, placing a greater burden on DCs to perform their task in less time while maintaining high accuracy. Automated classification algorithms may be able to aid DCs by improving accuracy or reducing time to review documents. However, because human DCs will likely never be replaced as the final decision-maker, and no algorithm will perform perfectly, it is crucial to understand the impact of automated classification suggestions on DC decision-making, especially algorithm errors. This will ensure that models incorporated into the DC workflow augment the analyst's capabilities without introducing unknown or unacceptable errors into the process.

Previous work has shown benefits to decision-making accuracy and efficiency when ML algorithms or other types of automated decision aids provide accurate information to analysts, in such fields as medical diagnostics (Wang & Summers, 2012), baggage screening (Rieger, Heilmann, & Manzey, 2021), object detection in overhead imagery (Kneusel & Mozer, 2017), visual search in both lab-based tasks and real-world imagery (e.g., Divis et al., 2021), and even identifying potential spam emails (Stites et al., 2021). As ML models improve in accuracy, incorrect model suggestions will become less frequent. Although this seems objectively good, humans are likely to miss rare events, a phenomenon known as the prevalence effect (Wolfe, Horowitz, & Kenner, 2005). In document classification detection, even one missed target (in this case, a classified document that a model failed to identify as classified) could result in the release of classified information, with potentially grave impacts for national security. It is thus critical to understand how DCs are impacted by both correct and incorrect model suggestions, to understand the risk/benefit trade-offs before implementation in an operational environment.

Much of the previous work investigating the use of ML decision aids has focused on target detection in visual imagery. The document classification domain is distinct in that it requires DCs to read documents and integrate this information with rules from a classification guide to identify sensitive information. This task requires extensive knowledge of the rules in the relevant guide(s) and technical jargon to identify whether a topic is present in a document. Failure to identify a document as classified poses the biggest risk to national security, although marking unclassified documents as being classified is also not desirable. Little work has investigated how people integrate ML suggestions into decisions about text. Lai and Tan (2019) found that an ML aid could help users identify whether text was deceptive or not. However, ML performance in their study was consistently high, and so it is not clear how participants recognized and overcame model errors.

In this study, we assessed the impact of automated classification decision algorithms on human decision-making in a simulated DC task. Because the task used differed from a typical DC workflow, we do not claim that our findings represent actual baseline performance. That being said, our study

takes an important first step in evaluating how an automated classification algorithm impacts DC performance relative to an experimentally determined baseline. We present our findings, along with recommendations for consideration before implementing such decision aids in a real DC workflow.

EXPERIMENT 1

In the first experiment, we assessed the human decision-making impact of two different types of automated classification algorithms: an ML algorithm and an ontological model. The details of the models are beyond the scope of paper, but an overview is as follows. The ML model was trained on a set of classified and unclassified documents in a particular subject area and identified key terms that differentiated them. Next, those terms were fed to multiple ML algorithms to produce an overall document score. For the ontological model, the relevant classification guide was used to create a model of relationships between concepts. Natural language processing techniques were then applied to a reviewed document and compared to the ontological model to identify rules from the guide. For both algorithms we predicted that, relative to baseline, correct model suggestions would improve decision accuracy, shorten response times, and increase user confidence.

METHOD

Fourteen participants took part in the study. All experimental protocols were approved by the Sandia National Laboratories Human Studies Board.

Materials consisted of 24 unique document excepts (16 classified, 8 unclassified) from a particular weapon system. The excerpts were approximately one paragraph (10 lines) in length, which did not differ statistically between classified and unclassified categories. Across both the classified and unclassified documents, 50% were shown with the correct classification suggestion, 25% with an incorrect suggestion, and 25% with no suggestion (baseline). Four experimental lists were created, and each participant completed one list. An incomplete Latin Squares counterbalancing design was used to ensure that each document appeared in a different condition in each list, to avoid effects of document content on results.

Stimuli were static screenshots of the documents displayed in a Sandia-developed web interface (see Figure 1). The document text was shown in

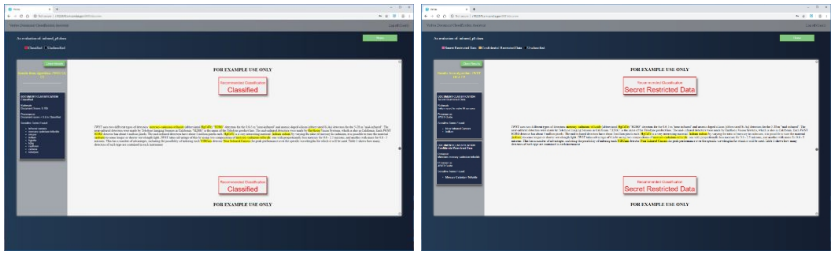

Figure 1: Experiment 1 stimuli examples for the ML (left) and ontological (right) conditions. The images presented here contain sample unclassified text.

the center/right portion of the screen, and the model output was shown in a sidebar on the left. The automated classification decision was shown at the top and bottom of the screen in colored text (red for *classified*, green for *unclassified*). A detailed description of the information displayed to users for each model output condition is listed in Table 1. Incorrect trials were experimenter-generated and maintained the same term list and highlighting as the correct decisions.

Table 1. Summary of conditions in Experiment 1. Note: Alg. = algorithm; ML = machine learning model; Ont. = ontological (rule-based) model; CG = classification guide.

Document Classification	Alg. Type	Model Correct		Model Incorrect
		Exact	Extra	Description
Classified	ML	Classified suggestion Term list Extra term highlighting Score >.5	Classified suggestion Term list Term highlighting Score >.5	Unclassified suggestion Term list Term highlighting Score <.5
	Ont.	Classification level & category Provenance (correct rules from CG) Term highlighting	Classification level & category Provenance (slightly wrong rules from CG) Term highlighting	Unclassified suggestion No rules from CG No term highlighting
	NA	No suggestion baseline		
Unclassified	ML	Unclassified suggestion Term list Term highlighting Score <.5		Classified suggestion Term list Term highlighting Score >.5
	Ont.	Unclassified suggestion No rules from CG No term highlighting		Classification level & category Provenance (wrong rules from CG) Term highlighting
	NA	No suggestion baseline		

Participants were instructed that although the documents would sometimes be accompanied by a suggested classification, they were responsible for the final classification determination. Participants were ensured that no documents would be marked or released based on their decisions. Documents were displayed to participants in a random order with a pre-determined model applied. Participants indicated their classification decision by clicking one of two buttons. Next, they indicated their confidence by clicking on a scale from 0–1 (where 0 = "Not at All Confident", .25 = "Slightly Confident", .50 = "Somewhat Confident", .75 = "Moderately Confident", and 1.00 = "Extremely Confident"). Response times were recorded for each trial. Experimental sessions lasted approximately 30 minutes.

RESULTS

Accuracy was measured using d', a target discriminability index. It was calculated by comparing the ratio between a participant's hit rate (i.e., classified documents accurately categorized as classified) and false alarm rate

(i.e., unclassified documents inaccurately categorized as classified). A d' score near 0 indicates chance performance; higher d' scores indicate better discriminability between document types. Response times and confidence scores were also analyzed.

D' scores were calculated for each participant, collapsing the different model conditions into three levels: correct suggestion, incorrect suggestion, and baseline. A one-way within-subjects Analysis of Variance (ANOVA) showed a main effect of algorithm correctness ($F_{(2,26)} = 33.75$, $p < .001$). Results are shown in Figure 2, and summarized in Table 2. Follow-up t-tests showed significant differences between all three conditions (all $t_{(13)} > 3.07$, $p < .01$). These results confirm our initial hypothesis that correct output from the algorithm would improve decision-making accuracy, and incorrect output would lower accuracy, relative to baseline.

Additional analyses tested the impact of the model correctness conditions (listed in Table 1) on participant accuracy, for classified documents only. Due to security concerns regarding the release of specific accuracy values, results will be discussed as a percentage change from baseline. Results from a one-way within-subjects ANOVA showed a main

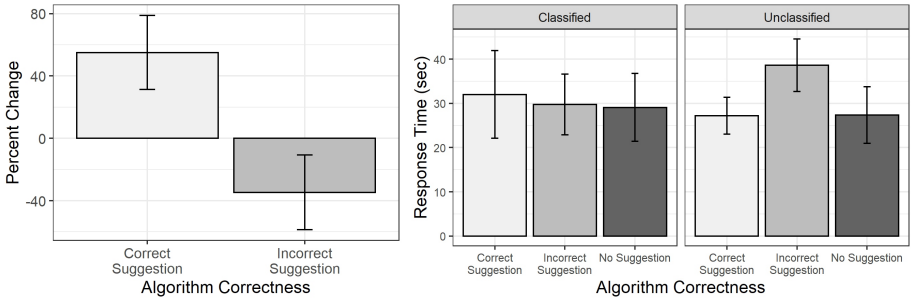

Figure 2: Experiment 1 results. Percent change in accuracy relative to baseline (left), and response times for accurate participant decisions (right). Error bars are 95% CIs.

Table 2. Summary of results for Experiments 1 and 2. Note: BL = baseline; asterisks (*) indicate a statistically significant difference; dashes (-) indicate no difference with baseline; E2 results indicate operationally relevant findings; NA indicates that condition did not exist.

Exp	Model Correctness	Participant Accuracy			Response Time (RT)		Confidence Score	
		Document Classification (C = Classified, U = Unclassified)						
		d' (*C* + *U*)	*C*	*U*	*C*	*U*	C	U
E1	Correct	* > BL	-	-	-	-	Correct > Incorrect	
	Incorrect	* < BL	ML-incor * < BL	-	-	* > BL		
E2	Correct	NA	> BL	-	< BL	-	-	-
	Incorrect	NA	-	NA	-	NA	Inacc. > Acc.	NA

effect of condition ($F_{(6,78)} = 4.86$, $p < .001$). Follow-up t-tests showed a significant 25% drop in accuracy for the ML-incorrect condition relative to baseline ($t_{(13)} = 2.75$, $p < .05$). No other effects were significant.

Next, the impact of algorithm correctness on response times (RTs) was assessed for accurate responses only. Due to missing data across conditions, RTs were analyzed using a linear mixed effects regression (LMER) model, with the fixed effect of algorithm correctness and random intercepts for each subject. Models were fit using the lme4 package (v. 1.1.19; Bates et al., 2010) and the afex package (v. 0.19.1) in R (v. 3.4.3). For classified documents, there was no significant effect of algorithm (*Chi-sq*(2) = 1.32, $p = .52$). For unclassified documents, there was a significant effect of algorithm correctness (*Chi-sq*(2) = 16.17, $p < .001$), driven by longer RTs to incorrect suggestions relative to baseline ($t = 3.53$, $p < .001$). In other words, people took longer to accurately identify unclassified documents when they were incorrectly labelled as classified.

There was a main effect of algorithm correctness on confidence scores ($F_{(2,26)} = 4.39$, $p < .05$), collapsing across classified and unclassified documents. Confidence was significantly higher for trials with a correct than incorrect suggestion ($t_{(13)} = 2.88$, $p < .05$). An identical pattern was observed for accurate responses only, but due to imbalanced trial numbers, the pairwise comparisons were not calculated.

DISCUSSION

Experiment 1 showed that, relative to baseline, correct algorithm suggestions significantly improved DCs' identification of classified information, and incorrect suggestions significantly decreased the identification of classified information. Findings suggest that a correct algorithm could *improve* DC accuracy by helping them identify classified information that they may have otherwise missed. On the other hand, an algorithm that fails to identify classified information could open the door to under-classification and the potential release of sensitive information.

There was a 25% drop in classification detection when the ML algorithm gave incorrect "Unclassified" suggestions to classified documents (accompanied by model explanations), but not when an incorrect decision was given without explanation. In other words, DCs were more likely to comply with an incorrect model suggestion when they received more information from the model. This result is consistent with previous work (Lai & Tan, 2019; Stites et al. 2021). Our findings raise an important potential security risk: DCs interpreted the provision of more information from the model as providing more evidence that classified information was present. Future work should carefully consider how the amount of model information shown to users will be interpreted in their risk assessment.

Response times were significantly longer than baseline when DCs accurately responded "Unclassified" to an unclassified document that the model wrongly suggested was classified. These longer response times were likely

caused by participants reading the paragraph closely to ensure no classified information was present. Although this is not a security risk per se, the benefits of classification identification versus the potential time impacts of erroneous suggestions should be weighed before the implementation of automated classification models. The fact that correct model suggestions did not shorten response times relative to baseline may have been an artifact of the short document length: the model suggestion conditions actually presented *more* information to participants than the baseline condition. It is possible that with longer documents, term highlighting will help DCs narrow in on the most critical areas of the document to direct their attention, whereas in Experiment 1, relative to the short content, reading this model-related information was time consuming. This prediction will be tested in Experiment 2.

EXPERIMENT 2

The goal of experiment two was to extend the findings from Experiment 1 into a more ecologically valid environment. To this end, full-length documents were used instead of single paragraph snippets, and DCs had access to an interactive classification guide. Only the ontological (rule-based) model was used, and the experiment-wise model accuracy was set to more closely reflect the known performance this model at the time of writing. We again tested the prediction that, relative to baseline, correct classification suggestions would improve participant decision-making accuracy, shorten response times, and increase confidence.

METHOD

Seven participants participated in the study; all protocols were approved by the Sandia Human Studies Board. Participants read up to 16 documents at their own pace (range: 5-14), 12 classified and four unclassified. Documents ranged in length from 1–113 pages and were drawn from the same weapon system as Experiment 1. For both classified and unclassified documents, 50% were shown with a model suggestion and 50% without. For classified documents, four (67%) were presented with correct suggestions, and two (33%) incorrect (wrong level). For unclassified documents, the suggestion was always correct (based on actual model performance). Counterbalancing ensured that documents shown with and without suggestions were rotated across participants. All participants saw the same first eight documents (six classified, two unclassified), in case participants did not finish in the allotted time. Because suggestions were generated by an existing model, documents shown with correct versus incorrect suggestions were different.

On each trial, participants saw the document and classification guide together in the same interface, with the document on the right and the guide on the left. Each document appeared with a document-level decision displayed at the top of the screen. The triggered rules from the classification guide were indicated with a checkmark and shown as annotations

(similar to document comments). Terms associated with each triggered rule were highlighted in the document. The classification guide was interactive: participants could click on annotations to jump to the text that triggered the rule. Users could also check/uncheck rules in the guide to dynamically change the level/category suggested by the model (this would not update the annotations or highlights, which were pre-loaded). For the baseline trials, no overall decision or term highlighting were displayed initially; clicking on the interactive guide could change the model-suggested document classification.

Participants read each document and provided their suggested document classification (level and category) along with their confidence (on a sliding scale from 0 = "Not confident at all" to 100 = "Completely confident"). Trial-level response times were collected, and number of clicks on the guide. At the end of the experimental session, participants completed two risk questionnaires; data from these assessments has been reported elsewhere (Fallon et al., 2021).

RESULTS

Due to the low number of participants and imbalanced trial counts across conditions, inferential statistics were not calculated. Instead, for each measure of interest, we present mean values with 95% confidence intervals around the mean (calculated using the $z = 1.96$) to estimate the size of condition-wise differences.

Trial-level accuracy was calculated for each participant; the appropriate classification level and category were required to be considered accurate. The percentage change in accuracy from baseline for correct and incorrect model suggestions was calculated next, separately for classified and unclassified documents. For classified documents, correct model suggestions improved participant accuracy by 100% (95% CI: [43, 157]) over baseline, whereas incorrect model suggestions improved participant accuracy 14% above baseline (95% CI: [-85, 113]; see Figure 3, panel A). Given that the incorrect suggestion's CI contains zero, we did not interpret this 14% difference as meaningful. Participant accuracy was identical for correct and incorrect suggestions to unclassified documents.

Response times for classified documents (accurate trials only) were shortest for correct model suggestions, with an average time savings of 42% relative to baseline (see Figure 3, panel B). Incorrect model suggestions produced RTs that were 10% faster than baseline, though with highly overlapping CIs. Response times for unclassified documents were almost identical for correct and no suggestion trials.

Confidence scores (for classified documents) were higher on trials that participants answered accurately than those they did not (see Figure 3, panel C). The exception was for classified documents for which participants agreed with an incorrect "Unclassified" suggestion. Confidence in these inaccurate decisions was higher than when participants accurately *disagreed* with the model's suggestion.

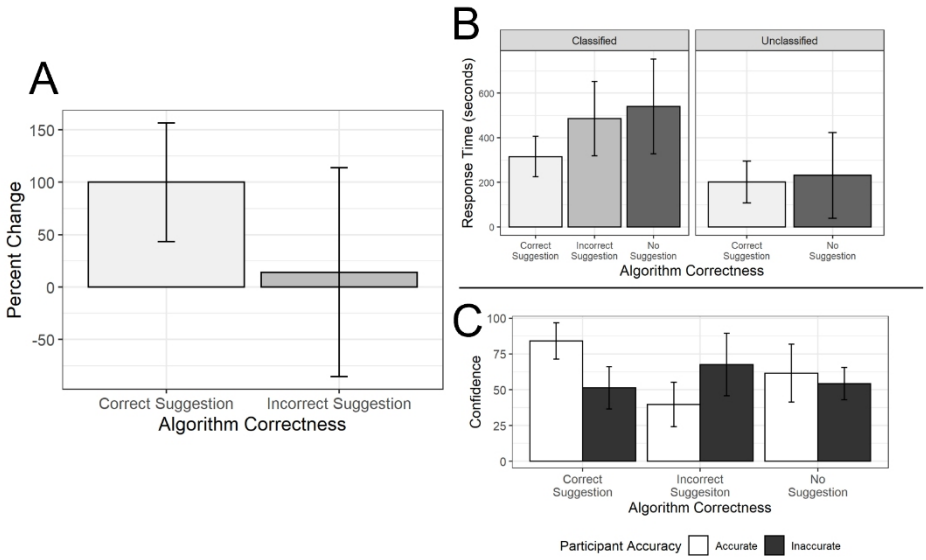

Figure 3: Experiment 2 results showing percent change in participant accuracy (classified documents, A), response times (accurate trials, B) and confidence (classified documents, C) based on algorithm correctness. Error bars are 95% CIs.

DISCUSSION

In Experiment 2, we found that correct model suggestions for classified documents increased decision accuracy, as in Experiment 1, and shortened response times relative to baseline. This finding suggests that interactive classification assistants could help DCs triage long documents to focus on the potentially sensitive sections, and ultimately make more accurate and faster decisions than the DC alone.

We also found that incorrect model suggestions for classified documents did not change accuracy or response times relative to baseline. This differed from Experiment 1, which showed lower discriminability for incorrect model decisions. It is possible that access to the interactive classification guide in Experiment 2 enabled participants to accurately respond despite model errors. However, participants reported higher confidence when they complied with an incorrect model suggestion for classified documents than when they (correctly) disagreed with the suggestion. These findings represent a potential security risk: the interactive model, in combination with high model accuracy, may make it more difficult for DCs to notice and overcome rare model errors.

CONCLUSION

Across two experiments, we assessed the impact of automated document classification suggestions on human decision-making accuracy, efficiency, and confidence. Our results consistently showed that correct model suggestions improved DC decision-making accuracy, and shortened response times for full-length documents. Findings suggest that these models could lighten the

load placed on DCs and improve the security posture around information release. However, our findings also highlight a few critical risks. When the model missed the identification of classified, DCs were also likely to miss it if an ML model provided extra information along with an incorrect suggestion. Additionally, when model errors were rare, DC confidence was high when they complied with such errors.

Our results should be interpreted with the following caveats. Our DC task (i.e., reading documents or paragraphs in isolation) is not representative of a real workflow. If DCs are unsure of their decisions, they can seek additional information by re-reading classification guides or consulting fellow DCs, rather than making a determination in the moment. DCs in our study were also unfamiliar with the model(s) used or how to incorporate the automated suggestions into their decision-making process. It is possible that with more exposure, they would learn the situations in which the model was reliable, or when to seek additional information. Future work should explore how model explanations, model error rate, and DC experience interact to impact DC performance. This work will be critical for recommending ways to incorporate automated classification suggestions into workflows to support DC decision-making, and ultimately, ensure the protection of sensitive information.

ACKNOWLEDGMENT

This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all right, title and interest in and to the article and is solely responsible for its contents. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this article or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan https://www.energy.gov/downloads/doe-public-access-plan. SAND2023-12489C.

REFERENCES

Bates, D. M. (2010). lme4: Mixed-effects modeling with R.

Divis, K., Howell, B., Matzen, L., Stites, M., & Gastelum, Z. (2021, December). The Cognitive Effects of Machine Learning Aid in Domain-Specific and Domain-General Tasks. In *Proceedings of the Annual Hawaii International Conference on System Sciences* (Vol. 2022).

Fallon, C. K., Brayfindley, E., Arneson, K., Brigantic, R., Stites, M., & Kittinger, L. (2021, October). A Methodology for Assessing Risk to Inform Technology Integration. In *2021 Resilience Week (RWS)* (pp. 1–7). IEEE.

Kneusel, R. T., & Mozer, M. C. (2017). Improving human-machine cooperative visual search with soft highlighting. *ACM Transactions on Applied Perception (TAP)*, *15*(1), 1–21.

Lai, V., & Tan, C. (2019, January). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. *In Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 29–38).

Rieger, T., Heilmann, L., & Manzey, D. (2021). Visual search behavior and performance in luggage screening: effects of time pressure, automation aid, and target expectancy. *Cognitive Research: Principles and Implications*, *6*(1), 1–12.

Stites, M. C., Nyre-Yu, M., Moss, B., Smutz, C., & Smith, M. R. (2021, July). Sage advice? The impacts of explanations for machine learning models on human decision-making in spam detection. In *International Conference on Human-Computer Interaction* (pp. 269–284). Springer, Cham.

Wang, S., & Summers, R. M. (2012). Machine learning and radiology. *Medical Image Analysis*, *16*(*5*), 933–951.

Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, *435*(7041), 439–440.