nature biomedical engineering

PEER-REVIEW INFORMATION

Detection of ovarian cancer via the spectral fingerprinting of quantum-defect-modified carbon nanotubes in serum by machine learning

Corresponding author: Daniel Heller

Editorial note

This document includes relevant written communications between the manuscript's corresponding author and the editor and reviewers of the manuscript during peer review. It includes decision letters relaying any editorial points and peer-review reports, and the authors' replies to these (under 'Rebuttal' headings). The editorial decisions are signed by the manuscript's handling editor, yet the editorial team and ultimately the journal's Chief Editor share responsibility for all decisions.

Any relevant documents attached to the decision letters are referred to as **Appendix #**, and can be found appended to this document. Any information deemed confidential has been redacted or removed. Earlier versions of the manuscript are not published, yet the originally submitted version may be available as a preprint. Because of editorial edits and changes during peer review, the published title of the paper and the title mentioned in below correspondence may differ.

Correspondence

Thu 01 Apr 2021
Decision on Presubmission Enquiry nBME-21-0747-PE

Dear Dr Heller,

Thank you for submitting to *Nature Biomedical Engineering* your Presubmission Enquiry, "Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer".

This is a compelling method for accurate disease detection (here, ovarian carcinoma) via the fingerprinting of a few relevant biomarkers in serum, by leveraging the sensitivity of modified carbon nanotubes and finding the best combination of biomarkers via machine learning. The performance comparisons and the classifier optimization make the manuscript stronger. We'll be happy to send the manuscript out for peer review when you formally submit it.

I should ask you to please fill in our reporting summary and policy checklist. (Please note that these forms are dynamic PDF files that can only be properly visualized and filled in by using Acrobat Reader.)

Both documents are aimed at ensuring good reporting standards and at easing the interpretation of results, and will be available to the reviewers. Should the manuscript be eventually published, the reporting summary will be attached to the published PDF of the paper and will also be available as supplementary information. More information is available on the editorial policies page.

Moreover, I recommend that you use the attached manuscript template. This will help you ensure that the manuscript complies with our data-presentation recommendations, that it includes all the necessary sections, and that it is structured to facilitate the assessment of peer reviewers and editors. In particular, please make sure that the manuscript provides thorough information on statistics and methods.

When you are ready to submit the manuscript, please upload the manuscript files as well as the reporting

summary and policy checklist.

Best wishes,

Pep

__
Pep Pàmies
Chief Editor, *Nature Biomedical Engineering*

Fri 28 May 2021
Decision on Article NBME-21-0747A

Dear Dr Heller,

Thank you again for submitting to *Nature Biomedical Engineering* your manuscript, "Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer", and for your patience in waiting for the feedback. The manuscript has been seen by three experts, yet one of them has yet to provide their report (we have chased them for weeks, and it is unclear whether they will provide a report). You will find two reviewer reports at the end of this message. I hope I will be able to provide you with additional feedback from a third reviewer in due course.

You will see that the two reviewers have some good words for the work, that they articulate concerns about the degree of support for the claims, and that in this regard they provide useful suggestions for improvement. We hope that with significant further work you can address the criticisms and convince the reviewers of the merits of the study. In particular, we would expect that a revised version of the manuscript provides:

* Evidence of performance under clinically relevant optimization criteria for the early detection of ovarian cancer, as argued by Reviewer #3.

* Given the diagnostic emphasis of the work, additional validation of the model, with an additional independent set of patient samples (rather than just using cross-validation).

When you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the (revised, if needed) reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

Please follow the following recommendations:

* Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).

* If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).

* If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).

* Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.

* The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers will the reports as they appear at the end of this message).

* Provide the rebuttal to the reviewer comments and the cover letter as separate files.

We hope that you will be able to resubmit the manuscript within 18 weeks from the receipt of this message. If this is the case, you will be protected against potential scooping. Otherwise, we will be happy to consider a revised manuscript as long as the significance of the work is not compromised by work published elsewhere or accepted for publication at *Nature Biomedical Engineering*.

We hope that you will find the referee reports helpful when revising the work. Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

—
Pep Pàmies

Chief Editor, *Nature Biomedical Engineering*

_____

Reviewer #2 (Report for the authors (Required)):

The authors presented a well written article for “Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer”. They report a machine model of a nanosensor array to collect physiochemical interactions to differentiate ovarian cancer from other diseases as well as healthy individuals. The report of sensitivity and selectivity are impressive. However, the reviewer has concerns about the nanosensor ability to recognize and differentiate high-grade serous ovarian carcinoma (HGSOC).

1. Could the authors elaborate on how and why these specific DNA sequences were used. As noted in the narrative, it is because they interact with the single walled carbon nanotubes (SWCNT). However, what is causing the sensor to recognize biological elements in the patient sample? It is not clear to the reader specifically how these sequences are used to differentiate samples particularly for HGSOC vs non-HGSCOC. For example, an antibody on a sensor surface can selectively bind to a protein. Do these DNA sequence act as aptamers? What is leading to this high selectivity?

2. It is stated that “to further assess the robustness of the sensor array and algorithm, we synthesized a new batch of OCC-DNAs.”, not clear here when you synthesized a new batch was this under the same conditions?

3. The authors suggest that unidentified biomarkers in serum are responsible for the sensor response as machine model detection of current ovarian serum markers resulted in high F-scores. This seems very interesting and also puzzling. Could the authors elaborate on their model to suggest why biomarker-dependent features scored highly? What are you detecting and how can it be validated through more traditional methods?

4. It is noted in support vector regression models that there was a prediction error in the normal concentration range for CA125 was larger than the high concentration range and it was attributed to the detection limit in single titration experiments. Can the authors elaborate on the detection limits that were selected in their model and how it relates to the clinically significant ranges for the known biomarkers?

5. Is this sensitivity and selectivity only from the perception-based model? If so, can the authors elaborate on the specifics of algorithm and what is unique about it? Also, the reviewer questions such a low sample size (269 patients) with this shotgun approach.

The work could advance some of the fundamental questions of evaluating current breast cancer biomarkers and possibly even biomarker panels if the authors can elaborate on their model specifically questions 3 and 5. However, the reviewer has concerns as to how this could relate to diagnostic testing. As to the number questions above question 1, 3, 4, and 5 are elements that the reviewer suggests are mandatory to address for further consideration in Nature Biomedical Engineering.

Reviewer #3 (Report for the authors (Required)):

Brief Summary
This manuscript describes organic color carbon nanotubes, an innovative technology for detecting physicochemical interactions with unknown components of serum from patients. The investigators trained this detection system to differentiate patients with ovarian cancer from patients with other conditions or no condition through analyses of the patients’ sera. Six artificial intelligence methods were applied with support vector machines (SVM) having the best performance with over-fitting ameliorated to the extent possible by 10-fold cross validation. The sera are from 269 female patients from six groups: 1. 56 high grade serous ovarian cancer (HGSOC) (most common, most lethal), 2. 71 from patients with normal conditions (although it’s unclear why they were patients), 3. 56 with other gynecologic diseases, 4. 61 with non-gynecologic diseases, 5. 7 in HGSOC remission, and 6. 18 in remission in other cancers. The 269 patients are split into a

training and validation set (n=215), and
a test set (n=54). The SVM approach produced the best result with 95% specificity compared to 81% for CA125, a known serum biomarker for ovarian cancer.

This manuscript describes a significant and original technological advance with carbon nanotubes with potential clinical applications to early detection of ovarian cancer via a blood test. If the test is significantly better than the current standard using CA125, then the impact on early detection of ovarian cancer could be substantial - and there is a clear need for such an improvement in the field. This technology could be applied across the whole field of diagnostic medicine, expanding beyond blood to urine, CSF, and other accessible bodily fluids with the potential of substantial improvements in all diagnostic areas.

This review focusses on the clinical application and does not address the technological advances in carbon nanotubes.

Major technical criticisms or questions:

1. Criterion for optimization should be appropriate for the clinical application. Early detection of ovarian cancer (and other cancers) requires sensitivity at a high specificity so that benefits outweigh harms. At most 10 surgeries to detect 1 ovarian cancer has been suggested as the limit for ovarian cancer screening. The current approach recently tested in a large screening
trial https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00731-5/fulltext uses a rising CA125 with referral to trans-vaginal sonography (TVS). It achieves just under 5 operations to detect 1 ovarian cancer, a 22% positive predictive value (PPV), and 84% of ovarian cancers detected prior to clinical detection (clinical sensitivity). To achieve this overall performance of the screening approach, the blood test has 98% specificity referring 2% of participants to TVS per year, and a sensitivity of 85%. The authors should compare the sensitivity of their blood test at the same specificity (98%) to
assess its potential for improving on a rising CA125.

2. The machine learning literature refers to clinical sensitivity as “recall” = true positives/(true positives + false negatives) = fraction of ovarian cancer cases detected by the test. The other parameter invoked is the precision = true positives/(true positives + false positives) which is the PPV in a clinical context. The authors minimize the F-score = 2/(1/recall + 1/precision) – the harmonic mean of the two parameters, and from a clinical context this criterion is equivalent to 2/(1/sensitivity + 1/PPV) which weights sensitivity and PPV equally. Machine learning has established methods for imbalanced learning so that sensitivity and PPV are not equally weighted. The authors should apply such methods to identify the nanotube test which maximizes sensitivity at 98% specificity = true negative / (true negative + false positive).

3. The abstract states that ovarian cancer is challenging to detect because biomarkers are elevated in other conditions, and that most biomarkers are not specific enough for screening purposes. This is not true. A rising CA125 followed by TVS for a positive blood test has a specificity of 99.8% and a PPV exceeding 20%, twice the minimum suggested in the literature. Instead, the challenge is to maximize sensitivity for early-stage disease while maintaining the specificity at an achievable 98%.

4. The abstract states that their strategy yields a specificity of 95% while CA125 has a specificity of 81%. These specificity values depend on which cutoff is chosen. To compare to the best method applied in clinical screening trials, a rising CA125 followed by TVS, the authors’ methods need to be compared to this first line blood test which has a specificity of 98% and a sensitivity of 84%.

5. Please explain the choice of other conditions since such conditions are rare in a general population screening context. If the aim is to develop a blood test for early detection of ovarian cancer, should not the “healthy controls” form the majority of the control group?

Minor technical questions:
1. Presumably cases were HGSOC while all other conditions were classified as “controls”. Please clarify.
2. The text states that “a single subsample was used to test the trained model”. Please clarify what testing in this context means, and why n=54 was chosen for the size of the test set.
3. The authors should define “recall” and “precision” and F-score for clinical readers in terms of true positives, false positives, true negatives, and false negatives, and relate these terms to “clinical sensitivity” and “positive predictive value”.

Missing citations to relevant literature
The authors should refer to the largest ovarian cancer screening trial:
1. https://pubmed.ncbi.nlm.nih.gov/26707054/, or
2. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00731-5/fulltext

Sat 05 Jun 2021
Decision on Article NBME-21-0747A

Dear Dr Heller,

Thank you once more for submitting to *Nature Biomedical Engineering* your manuscript, "Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer". I have now received an additional report (from Reviewer #1), which I include at the end of this message alongside the other two reports that I sent to you about a week ago.

Reviewer #1 raises criticisms, mostly focused on the performance of the sensor and on its potentially clinical applicability, that largely overlap with those of the other two reviewers. As I outlined in my previous message, I hope that with significant further work you can address the criticisms and convince the reviewers of the merits of the study. In particular, we would expect that a revised version of the manuscript provides:

* Evidence of performance under clinically relevant optimization criteria for the early detection of ovarian cancer, as argued by Reviewer #3.

* Given the diagnostic emphasis of the work, additional validation of the model, with an additional independent set of patient samples (rather than just using cross-validation).

* Description of the method for choosing the relevant oligonucleotide sequences for the assay, and further insight into why using only biomarker-dependent features in the model did not result in optimal performance.

When you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the (revised, if needed) reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

Please follow the following recommendations:

* Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).

* If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).

* If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).

* Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.

* The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers will the reports as they appear at the end of this message).

* Provide the rebuttal to the reviewer comments and the cover letter as separate files.

We hope that you will be able to resubmit the manuscript within 18 weeks from the receipt of this message. If this is the case, you will be protected against potential scooping. Otherwise, we will be happy to consider a revised manuscript as long as the significance of the work is not compromised by work published elsewhere or accepted for publication at *Nature Biomedical Engineering*. Because of the COVID-19 pandemic, should you be unable to carry out experimental work in the near future we advise that you reply to this message with a revision plan in the form of a preliminary point-by-point rebuttal to the comments from all reviewers that also includes a response to any points highlighted in this decision. We should then be able to provide you with additional feedback.

We hope that you will find the referee reports helpful when revising the work. Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

__
Pep Pàmies
Chief Editor, *Nature Biomedical Engineering*

_____

Reviewer #1 (Report for the authors (Required)):

In this manuscript, the authors reported a nanosensor array composed of multiple ssDNA-encapsulated organic color center-modified carbon nanotubes (OCC-DNAs) together with a machine learning model to achieve perception-based detection of ovarian cancer from patient serum samples. The optimal model in this study yielded 95% specificity, which outperformed the serum biomarker cancer antigen with a specificity of only 81% (CA125). Most importantly, the presented technology required a small sample volume from each patient (360 μL) and relatively short testing time (~ 2 hrs), potentially deriving benefits for clinical screening of ovarian cancers. However, the authors did not explain the underlying mechanisms of the proposed nanosensor array and the specific function of encapsulating ssDNA. Moreover, the sample size was relatively small for an established and reliable prediction model. The authors should discuss more on the biases, limitations, robustness, and generalizability of
their model. The following issues also need to be addressed.

1. Regarding the design of the study, what is the standard of ssDNA selection for SWNT encapsulation? The authors only mentioned that the ssDNA was chosen based on their ability to form tight structures on the SWCNT surface; however, no explanations were provided regarding their interactions with potential biomarkers of ovarian cancers and the underlying mechanisms behind the selection. Moreover, it seems that the proposed nanosensor was not specific towards ovarian cancer, then how can the proposed achieve accurate screening of ovarian cancer?

2. In some recent studies, machine learning has been deployed to facilitate multi-analyte screening test combing both ctDNA and protein biomarker tests (Science 2018, 359, 926; Science 2020, 369, eabb9601), which can achieve >99% specificity and >95% sensitivity for ovarian cancer, as well as identification of other types of cancers. Is the presented model cost-effective enough and amenable to high throughput? What are the advantages compared to these methods that appear simpler?

3. The sample size of the dataset was 269, including training, validation, and test sets, which was quite small to establish an algorithm with good generalizability in a larger population. Moreover, the authors should also specify how the data were assigned into these partitions and whether there are any systemic differences between the data in each partition.

4. The authors classified the patient population into three categories – HGSOC, other non-HGSOC diseases, and healthy individuals. From the biological point of view, there may be a wide discrepancy in the test results among the non-HGSOC disease category, especially when low-grade ovarian carcinoma and some other cancers are included. The generalization of the presented model is doubtful when applying to real scenario screening.

5. Although the authors pointed out the potential of the technology for clinical translation and biomarker discovery, the authors did not mention 1) how the proposed technology can improve the performance and efficiency of clinicians; 2) how the feature values influence the prediction outcomes; 3) and how these features can link to the underlying biology of ovarian cancer.

6. The authors should compare the performance between the presented model to previously published models to show its superiority. Moreover, apart from the comparison of specificity between the presented model with the CA125 test, the estimated screening cost should also be compared.

7. The risk of bias should be assessed in terms of participants, predictors, outcomes, and analysis.

8. According to Figure 2b and Figure S2, the spectral differences of E11 and E11' between HGSOC and benign samples were only within 2 nm in many groups. Such a slight difference was not sufficient to distinguish the pathological status because many factors (e.g., sample preparation, ion concentration, etc.) can cause small errors in spectral measurements.

9. Following the previous question, the specificity of such sensor to detect cancer under the backdrop of other diseases is doubted. For instance, to detect ovarian cancer in aged patient or patient with autoimmune diseases, such as diabetes? These background chronic diseases can affect the blood constituents. The authors need to validate this.

10. The authors only validated the performance using the test data from the same source. It is better to include both internal and external validation to provide more insight into the generalization of the model.

11. In Figure S3, "a", "b", "c" and "d" appeared in the figure caption; however, they can no be found the Figure. The authors are recommended to check the figures carefully to make sure there is no missing or errors in each figure.

12. The authors claimed that there is linear correlation of each sensor response to medications in Figure S5. However, we did not find any linear correlation in Figure S5. How did the authors make this conclusion?

Reviewer #2 (Report for the authors (Required)):

The authors presented a well written article for "Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer". They report a machine model of a nanosensor array to collect physiochemical interactions to differentiate ovarian cancer from other diseases as well as healthy individuals. The report of sensitivity and selectivity are impressive. However, the reviewer has concerns about the nanosensor ability to recognize and differentiate high-grade serous ovarian carcinoma (HGSOC).

1. Could the authors elaborate on how and why these specific DNA sequences were used. As noted in the narrative, it is because they interact with the single walled carbon nanotubes (SWCNT). However, what is causing the sensor to recognize biological elements in the patient sample? It is not clear to the reader specifically how these sequences are used to differentiate samples particularly for HGSOC vs non-HGSCOC. For example, an antibody on a sensor surface can selectively bind to a protein. Do these DNA sequence act as aptamers? What is leading to this high selectivity?

2. It is stated that "to further assess the robustness of the sensor array and algorithm, we synthesized a new batch of OCC-DNAs.", not clear here when you synthesized a new batch was this under the same conditions?

3. The authors suggest that unidentified biomarkers in serum are responsible for the sensor response as machine model detection of current ovarian serum markers resulted in high F-scores. This seems very interesting and also puzzling. Could the authors elaborate on their model to suggest why biomarker-dependent features scored highly? What are you detecting and how can it be validated through more traditional methods?

4. It is noted in support vector regression models that there was a prediction error in the normal concentration range for CA125 was larger than the high concentration range and it was attributed to the detection limit in single titration experiments. Can the authors elaborate on the detection limits that were selected in their model and how it relates to the clinically significant ranges for the known biomarkers?

5. Is this sensitivity and selectivity only from the perception-based model? If so, can the authors elaborate on the specifics of algorithm and what is unique about it? Also, the reviewer questions such a low sample size (269 patients) with this shotgun approach.

The work could advance some of the fundamental questions of evaluating current breast cancer biomarkers and possibly even biomarker panels if the authors can elaborate on their model specifically questions 3 and 5. However, the reviewer has concerns as to how this could relate to diagnostic testing. As to the number

questions above question 1, 3, 4, and 5 are elements that the reviewer suggests are mandatory to address for further consideration in Nature Biomedical Engineering.

Reviewer #3 (Report for the authors (Required)):

Brief Summary
This manuscript describes organic color carbon nanotubes, an innovative technology for detecting physicochemical interactions with unknown components of serum from patients. The investigators trained this detection system to differentiate patients with ovarian cancer from patients with other conditions or no condition through analyses of the patients' sera. Six artificial intelligence methods were applied with support vector machines (SVM) having the best performance with over-fitting ameliorated to the extent possible by 10-fold cross validation. The sera are from 269 female patients from six groups: 1. 56 high grade serous ovarian cancer (HGSOC) (most common, most lethal), 2. 71 from patients with normal conditions (although it's unclear why they were patients), 3. 56 with other gynecologic diseases, 4. 61 with non-gynecologic diseases, 5. 7 in HGSOC remission, and 6. 18 in remission in other cancers. The 269 patients are split into a training and validation set (n=215), and
a test set (n=54). The SVM approach produced the best result with 95% specificity compared to 81% for CA125, a known serum biomarker for ovarian cancer.

This manuscript describes a significant and original technological advance with carbon nanotubes with potential clinical applications to early detection of ovarian cancer via a blood test. If the test is significantly better than the current standard using CA125, then the impact on early detection of ovarian cancer could be substantial - and there is a clear need for such an improvement in the field. This technology could be applied across the whole field of diagnostic medicine, expanding beyond blood to urine, CSF, and other accessible bodily fluids with the potential of substantial improvements in all diagnostic areas.

This review focusses on the clinical application and does not address the technological advances in carbon nanotubes.

Major technical criticisms or questions:

1. Criterion for optimization should be appropriate for the clinical application. Early detection of ovarian cancer (and other cancers) requires sensitivity at a high specificity so that benefits outweigh harms. At most 10 surgeries to detect 1 ovarian cancer has been suggested as the limit for ovarian cancer screening. The current approach recently tested in a large screening
trial https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00731-5/fulltext uses a rising CA125 with referral to trans-vaginal sonography (TVS). It achieves just under 5 operations to detect 1 ovarian cancer, a 22% positive predictive value (PPV), and 84% of ovarian cancers detected prior to clinical detection (clinical sensitivity). To achieve this overall performance of the screening approach, the blood test has 98% specificity referring 2% of participants to TVS per year, and a sensitivity of 85%. The authors should compare the sensitivity of their blood test at the same specificity (98%) to
assess its potential for improving on a rising CA125.

2. The machine learning literature refers to clinical sensitivity as "recall" = true positives/(true positives + false negatives) = fraction of ovarian cancer cases detected by the test. The other parameter invoked is the precision = true positives/(true positives + false positives) which is the PPV in a clinical context. The authors minimize the F-score = 2/(1/recall + 1/precision) – the harmonic mean of the two parameters, and from a clinical context this criterion is equivalent to 2/(1/sensitivity + 1/PPV) which weights sensitivity and PPV equally. Machine learning has established methods for imbalanced learning so that sensitivity and PPV are not equally weighted. The authors should apply such methods to identify the nanotube test which maximizes sensitivity at 98% specificity = true negative / (true negative + false positive).

3. The abstract states that ovarian cancer is challenging to detect because biomarkers are elevated in other conditions, and that most biomarkers are not specific enough for screening purposes. This is not true. A rising CA125 followed by TVS for a positive blood test has a specificity of 99.8% and a PPV exceeding 20%, twice the minimum suggested in the literature. Instead, the challenge is to maximize sensitivity for early-stage disease while maintaining the specificity at an achievable 98%.

4. The abstract states that their strategy yields a specificity of 95% while CA125 has a specificity of 81%. These specificity values depend on which cutoff is chosen. To compare to the best method applied in clinical screening trials, a rising CA125 followed by TVS, the authors' methods need to be compared to this first line blood test which has a specificity of 98% and a sensitivity of 84%.

5. Please explain the choice of other conditions since such conditions are rare in a general population screening context. If the aim is to develop a blood test for early detection of ovarian cancer, should not the "healthy controls" form the majority of the control group?

Minor technical questions:
1. Presumably cases were HGSOC while all other conditions were classified as "controls". Please clarify.
2. The text states that "a single subsample was used to test the trained model". Please clarify what testing in this context means, and why n=54 was chosen for the size of the test set.
3. The authors should define "recall" and "precision" and F-score for clinical readers in terms of true positives, false positives, true negatives, and false negatives, and relate these terms to "clinical sensitivity" and "positive predictive value".

Missing citations to relevant literature
The authors should refer to the largest ovarian cancer screening trial:
1. https://pubmed.ncbi.nlm.nih.gov/26707054/, or
2. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00731-5/fulltext

Tue 21 Dec 2021
Decision on Article NBME-21-0747B

Dear Dr Heller,

Thank you for your patience in waiting for all the reports on your revised manuscript, "Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer". As noted in previous e-mail correspondence, the manuscript has been seen by the original reviewers. In their reports, which you will find at the end of this message, you will see that Reviewers #1 and #3 raise a few more technical points that I am sure you will be able to address.

As before, when you are ready to resubmit your manuscript, please upload the revised files, a point-by-point rebuttal to the comments from all reviewers, the reporting summary, and a cover letter that explains the main improvements included in the revision and responds to any points highlighted in this decision.

As a reminder, please follow the following recommendations:

* Clearly highlight any amendments to the text and figures to help the reviewers and editors find and understand the changes (yet keep in mind that excessive marking can hinder readability).

* If you and your co-authors disagree with a criticism, provide the arguments to the reviewer (optionally, indicate the relevant points in the cover letter).

* If a criticism or suggestion is not addressed, please indicate so in the rebuttal to the reviewer comments and explain the reason(s).

* Consider including responses to any criticisms raised by more than one reviewer at the beginning of the rebuttal, in a section addressed to all reviewers.

* The rebuttal should include the reviewer comments in point-by-point format (please note that we provide all reviewers will the reports as they appear at the end of this message).

* Provide the rebuttal to the reviewer comments and the cover letter as separate files.

I look forward to receive a further revised version of the work. Please do not hesitate to contact me should you have any questions.

With best wishes for the holidays,

Pep

__
Pep Pàmies
Chief Editor, *Nature Biomedical Engineering*

_____
Reviewer #1 (Report for the authors (Required)):

In this revised manuscript, the authors have made their effort to address the issues and comments raised by the reviewers. The quality of this manuscript has been improved after the substantial revisions and appended discussions on the design rationale, superiorities over the conventional tests, and bias analysis. Although the underlying interactions and types of proteins that contribute to the signal changes of the sensors remain unclear, it may be investigated using proteome analysis in the future to benefit pathophysiology study. However, the following minor concerns need to be addressed before the acceptance of manuscript.

1. The authors mentioned that the proposed sensors have non-stringent storage requirements and low

screening cost compared to CA125 blood test. Then, what is the shelf life of these DNA-wrapped SWCNTs? What are the required storage conditions of these SWCNTs? If stored at ambient conditions, how long can DNA coating maintain its integrity and produce reproducible results? The authors should provide data on this. I think that the authors should tone down on the cost statement.

2. How is the reproducibility of the properties and test results of OCC-DNAs from different batches? It is known that wrapped SWCNT might differ in bioaffinity from batch to batch.

3. How good is the correlation of the test results with the staging of ovarian cancer?

4. What is the earliest cancer stage that the proposed in your sample population?

Reviewer #2 (Report for the authors (Required)):

Reviewer 2 notes the improvements the authors made to the manuscript and suggest for publication in Nature Biomedical Engineering.

Reviewer #3 (Report for the authors (Required)):

Major issues

#1. The authors use an estimate of 87% sensitivity at 98% specificity. However, given the results in Fig. 3, the more appropriate and robust estimate is 77% sensitivity at 98% specificity, achieved by weighting the PPV more than the sensitivity (weight ß = 0.2). The authors do make a caveat correctly that their markers achieve this sensitivity in clinically diagnosed patients, while the screening studies (UKCTOCS) are in asymptomatic patients where sensitivity is expected to be significantly lower.

#2. In response to my recommendation (#2) to maximize sensitivity at 98% specificity by using a weighted harmonic mean. They argue that weights less than 1.0 weigh false positives more importantly than true positives. They investigated weights from 3 down to 0.2, with the smaller weights systematically increasing sensitivity at 98% specificity, although this increase was statistically insignificant. However, the median sensitivity changes from a low of 25% to a high of 77%. This change is of great clinical significance, and it's not clear what test was used to conclude the change was not statistically significant. I am not clear why the authors conclude that their method yields a sensitivity of 87% at 98% specificity when ß is 1.0 since the graph (Fig. 3 under "Page 5") clearly shows the optimal sensitivity is 77% when ß is 0.2. The authors continue to claim 87% sensitivity at 98% specificity. Fig. 3 shows that 77% is the more appropriate estimate of sensitivity at 98%
specificity, and this sensitivity estimate should be the comparison with the UKCTOCS results of 84% sensitivity at 98% specificity.

The issue of specificity for ovarian cancer screening has been addressed and overcome by Jacobs' series of screening studies showing that a multimodal screen with a blood test (CA125) followed by TVS for patients with a positive blood test yields an acceptable PPV and thus is sufficiently specific for screening. This series of studies stands in contrast to the authors' assertion in the Abstract that biomarkers are not specific for ovarian cancer. The challenge for ovarian cancer biomarkers as a first line test leading to imaging as a second line test is not specificity, it is that the initial test is not sensitive enough to detect a high proportion of cases in early stage. This challenge of low sensitivity for early-stage disease should be the clear statement of the clinical challenge facing biomarker researchers in identifying a first line test for early detection of ovarian cancer.

#3. The authors assert in the Abstract that the markers lack specificity because of the elevation in other conditions. This is also incorrect in that even though CA125 is elevated in other conditions, CA125 does not lack specificity when followed with TVS for positive CA125 tests. The elevation in these other conditions is sufficiently rare that a high specificity can be achieved despite their (infrequent) presence. In the vast majority of cases where CA125 is elevated when screening, no other conditions are identified, and subsequently the CA125 usually returns to baseline or stops increasing and becomes stable over time.

In Page 1, the authors state that “screening strategies are associated with a high rate of false-positive results and a risk of harm from invasive testing”. For UKCTOCS, there was not a high rate of false positive results. The literature had set an upper limit of 10 operations to identify 1 ovarian cancer, and UKCTOCS multimodal arm achieved a rate of 5 operations to identify 1 ovarian cancer, half the rate stated in the literature and therefore within acceptable limits (PPV > 10%). The authors should emphasize that the challenge is increasing the sensitivity for detection of early-stage disease while maintaining the already achieved high specificity (98% for blood test, 99.8% for blood test followed by TVS) and high PPV > 20%.

#4. The new markers’ sensitivity estimate of 87% at 98% specificity should be 77% by Fig. 3.

#5. The authors change the clinical challenge from early detection of ovarian cancer in the asymptomatic population to differentiating ovarian cancer from patients with other diseases. They should be consistent. A better approach would be to argue that they are developing the markers to differentiate ovarian cancer from other diseases as a first step, and that in future studies, they will examine the performance of their markers in the context of screening, where samples from healthy individuals will form the bulk of the control group.

Mon 10 Jan 2021
Decision on Article NBME-21-0747C

Dear Dr Heller,

Thank you again for the latest version of your manuscript, "Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer". You will see in the reports included below that Reviewer #3 has a few more recommendations for caveats that would need to be added to the text so as to quality the statements about the performance of the assay and the specifics of its clinical applicability.

As before, when you have amended the manuscript and responded to the reviewer's points, please upload the revised files. I will then carry out a round of editorial checks to help you bring the manuscript in final shape.

Best wishes,

Pep

__
Pep Pàmies
Chief Editor, *Nature Biomedical Engineering*

_____
Reviewer #1 (Report for the authors (Required)):

The authors have made their effort to properly address my concerns and it is recommended for publication.

Reviewer #3 (Report for the authors (Required)):

Review of “Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer”.
9 January 2022
Reviewer #3

#1. The authors clarify that their estimate of 87% sensitivity at 98% specificity is the maximum performance over “all possible combinations of sensor array combinations, composed of up to 6 OCC-DNAs at 98% specificity”. They helpfully provide the red line in Fig 3d which shows the sensitivity for the optimal combination. Because it is the optimal combination chosen from many combinations, this level of performance is unlikely to validate due to some component of the performance resulting from random chance variation even in the context of cross validation. How much is due to chance is difficult to evaluate in these high-dimensional situations, but due to this effect of random variation the estimate of 87% sensitivity is optimistically upwardly biased, and due to the apparent large number of combinations this bias can be substantial. Because of this intractable mathematical challenge of reducing the estimate by the appropriate amount, the best approach to obtaining an unbiased
estimate of performance is to validate the optimal marker combination in an independent sample set of cases and controls, often from different sources than the discovery sample set. Cross validation can reduce the bias but it usually does not remove the bias completely. Samples from ovarian cancer patients, especially early stage patients, are difficult to obtain due to its very low incidence. It is likely that the authors’ best approach to address this issue is to acknowledge it with a caveat since I doubt they have access to an independent validation sample set. They should state that the results they present are in the context of samples obtained at diagnosis usually due to symptoms, they are likely upwardly biased due to choosing the maximum across a large number of candidate combinations (and state how many combinations were evaluated), but they are to be viewed as a stepping stone to further work in samples obtained prior to diagnosis usually obtained during the conduct of
screening trials and from which stronger claims about earlier detection can be substantiated.

#2. I am not convinced that ANOVA (with a multiple comparisons adjustment) is the best test for assessing whether the sensitivity has increased when ß is changed. There would appear to be multiple dependencies between the estimates of sensitivity at different ß whereas ANOVA assumes independence of the values at different ß. However this is not a critical point in the manuscript.
My preference for the statement in the abstract is to be clear about the clinical challenge in early detection of ovarian cancer rather than to use the vague term “not [] sufficient diagnostic power”. A clear statement would help the field focus on the appropriate challenge, namely lack of sensitivity for early stage disease, rather than the vague challenge of insufficient power. “Serum biomarkers are widely used as diagnostic indicators, but many are not sensitive enough for early stage disease when used for screening. Ovarian cancer, for example, is challenging to diagnose, in part because the few established biomarkers for the disease, while providing sufficient specificity when followed by ultrasound, do not provide sufficient sensitivity for early stage disease.”

#3. The Page 1 conclusion would be clearer with the addition of the underline phrase: but the benefit of screening is still elusive due to lack of sensitivity for early stage disease.

#4. The authors have explained their estimate of 87% in the latest revision. However they need to be cautious and acknowledge an unknowable upward bias, since the estimate is the maximum across a large number of candidate combinations albeit cross-validated estimates.

#5. The authors have changed the language but only partly address this concern. In my judgment, due to the upward bias of the estimate due to the choice of the maximum across many candidates, there is insufficient evidence to assert that the performance exceeds current best clinical screening test (87% vs 84% sensitivity). I would accept an assertion that the performance is similar to the current best clinical screening test.
It is widely acknowledged that it is easier to “detect” symptomatic cases compared to controls, as the authors do in this study, than to detect asymptomatic cases amongst an apparently healthy population (screening). Therefore this sentence should be modified as underlined: However, considering the fact that specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, performance is likely to be lower in the clinical screening setting in which specimens are obtained in asymptomatic individuals before clinical diagnosis.

Sun 06 Feb 2021
Decision on Article NBME-21-0747D

Dear Dr Heller,

Thank you for your revised manuscript, "Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer". As noted in previous e-mail correspondence, having consulted with Reviewer #3 (whose comments you will find attached), I am pleased to write that we shall be happy to publish the manuscript in *Nature Biomedical Engineering*, provided that the suggestions of the reviewer are considered and that the points specified in the attached instructions file are addressed.

When you are ready to submit the final version of your manuscript, please upload the files specified in the instructions file.

For primary research originally submitted after December 1, 2019, we encourage authors to take up transparent peer review. If you are eligible and opt in to transparent peer review, we will publish, as a single supplementary file, all the reviewer comments for all the versions of the manuscript, your rebuttal letters, and the editorial decision letters. **If you opt in to transparent peer review, in the attached file please tick the box 'I wish to participate in transparent peer review'; if you prefer not to, please tick 'I do NOT wish to participate in transparent peer review'.** In the interest of confidentiality, we allow redactions to the rebuttal letters and to the reviewer comments. If you are concerned about the release of confidential data, please indicate what specific information you would like to have removed; we cannot incorporate redactions for any other reasons.
More information on transparent peer review is available.

Please do not hesitate to contact me should you have any questions.

Best wishes,

Pep

__
Pep Pàmies
Chief Editor, *Nature Biomedical Engineering*

P.S. Nature Research journals encourage authors to share their step-by-step experimental protocols on a protocol-sharing platform of their choice. Nature Research's Protocol Exchange is a free-to-use and open resource for protocols; protocols deposited in Protocol Exchange are citable and can be linked from the published article. More details can be found at www.nature.com/protocolexchange/about.

_____
Reviewer #3 (Report for the authors (Required)):

Please see the attached file.

Appendix #1

_____
Nature Biomedical Engineering is a Transformative Journal. Authors may publish their research with us through the traditional subscription access route, or make their paper immediately open access through payment of an article-processing charge. More information about publication options is available.

You may need to take specific actions to comply with funder and institutional open-access mandates. If the work described in the accepted manuscript is supported by a funder that requires immediate open access

(as outlined, for example, by Plan S) and your manuscript was originally submitted on or after January 1st 2021, then you will need to select the gold OA route. Authors selecting subscription publication will need to accept our standard licensing terms (including our self-archiving policies), and these will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Review of "Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer".

26 January 2022

Reviewer #3

The authors state in the abstract that their strategy yielded 87% sensitivity at 98% specificity, and then compare this result to a large scale randomized early detection trial spanning 15 years in over 200,000 women. However, their sensitivity/specificity combination is tested in symptomatic patients at diagnosis which will have a much greater sensitivity than when the same test is performed in asymptomatic patients in a screening. So, this is an "apples to oranges" comparison, and unfairly compares their classifier's performance in symptomatic patients to the results of another test applied to screening asymptomatic individuals. Further, since this is the performance in the training set of the best of many candidate classifiers, there is an upward bias in choosing the best classifier. The choice of the best performing classifier and quoting in the abstract its performance in the training set is akin to sampling many points from a standard Normal distribution with mean 0 (and SD=1 say) and selecting the largest. The largest will almost always be positive, it will not have an expected value of 0, it will be biased upwards. Similarly, their choice of the best performing classifier is upwardly biased. In an independent validation set it will have a lower performance, that is, a lower sensitivity, or lower specificity, or most likely both lower sensitivity and lower specificity. The authors then apply the classifier to an independent validation sample set with 7 HGSOC cases and 47 controls where their results have an AUC = 1, that is, perfect classification. Even when cross-validation was applied, the AUC was still 0.97, a very high value. They do not quote the AUC =1 (sensitivity = 100%, specificity = 100%) result in the abstract, probably because it would raise too many red flags in the minds of readers. Having 7 cases in a validation sample set is overwhelmingly insufficient to validate the claim of high sensitivity (87%) at a high specificity of 98%.

> **Page 5**: "The best performing prediction model was the sensor array combination of 4-N$(C_2H_5)_2$*C$T_2C_3T_2$C, 4-N$(C_2H_5)_2$*$(TAT)_4$, 3,4,5-F_3*$(TAT)_4$, 3,4,5-F_3*$(AT)_{15}$, and 3,4,5- F_3*$(GT)_{15}$, and yielded 87% sensitivity at 98% specificity when PPV and sensitivity were equally weighted (β =1). "
>
> **Abstract**: This strategy yielded 87% sensitivity at 98% specificity versus 84% via the multimodal test using the biomarker cancer antigen 125 and transvaginal ultrasonography.
>
> **Page 5:** These values are consistent with the cross- validation scores and gave a similar receiver operating characteristic (ROC) curve (Fig 3e), indicating that the model did not overfit the data.

In their rebuttal, the authors claim that "However, throughout the model optimization processes, there is no evidence and/or artifacts that would lead to upwardly biased results". The choice of the optimum classifier of many classifiers is the evidence that would lead to upwardly biased results. The estimate of 87% sensitivity is upwardly biased, it will be lower in an independent validation set. And the validation set would require many more cases than the authors' validation set with only 7 HGSOC. This small number of cases has little power to validate.

In their revision, the acknowledgment of the difference between their setting of clinically diagnosed patients and screening asymptomatic individuals is vague: "…. prediction outcomes may differ ….". The authors should be clear that their classifier will have lower performance when applied to a screening setting than when applied to clinically diagnosed symptomatic patients.

I agree with the authors' rebuttal that this technology can be applied to other settings than early detection as stated in their rebuttal: "We agree that early-stage detection is an important challenge in the area of ovarian cancer detection but not the only challenge." However, when applying their technology to early detection, the authors need to be clear what the challenge is. Achieving a sufficiently high specificity so that the PPV is acceptable has been solved and the authors should state that it has been solved. What has not been solved is sufficient sensitivity for early stage disease while maintaining the already achieved high specificity and PPV. That is the challenge for early detection of ovarian cancer and the authors need to make both of these points clear.

A related part of their rebuttal states: "Therefore, we believe that it is not appropriate to introduce this work focused solely on early-stage detection as the only clinical use of this technology/platform in the abstract." This reviewer is not requesting that the work focus solely on early stage detection. I am requesting that when the authors raise the topic of early detection (screening), that the challenge for early detection be clearly stated as a lack of sensitivity for early stage disease (which includes STICs as part of the disease process to be detected earlier than when symptoms arise) while maintaining the already achieved sufficiently high specificity. For example, the abstract could be improved with the underlined words as follows:

> In Abstract:
> "Serum biomarkers are widely used in medical decision making, but in many cases are not sensitive and/or specific enough to be diagnostic, especially in cases of low volume disease. In the quest to develop a screening test for ovarian cancer, for example, the few established biomarkers while achieving sufficient specificity do not enable sufficient sensitivity of patients with early-stage disease to impact mortality."

From my previous review:

> #3. The Page 1 conclusion would be clearer with the addition of the underline phrase: but the benefit of screening is still elusive due to lack of sensitivity for early stage disease.

The authors' rebuttal is: Addressed above in comment #2.

My critique is not addressed in comment #2. When the authors choose to introduce screening, they should be clear what the challenge for screening is. Instead, they state the application of the technology is broader than early detection. I agree that is the case, but if the authors introduce screening, then be clear what the challenge for screening for ovarian cancer is, namely lack of sensitivity for early stage disease.

The authors' rebuttal: This is relevant to Comment #1. We understand the reviewer's concern and thus we modified the discussion as follows:

> In Page 8,
> "... specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, prediction outcomes may differ in the clinical screening setting in which specimens are obtained in asymptomatic individuals before clinical diagnosis. Further studies, in the context of screening the general population, are warranted to evaluate the ability of the technology to identify pre-invasive and early-invasive disease."

This "may differ" is not sufficiently clear. It is widely acknowledged that it is easier to "detect" symptomatic cases compared to controls, as the authors do in this study, than to detect asymptomatic cases amongst an apparently healthy population (screening). Therefore this sentence should be modified as underlined:

In Page 8,
“... specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, performance will be lower in the screening setting in which specimens are obtained in asymptomatic individuals before clinical diagnosis. Further studies, in the context of screening the general population, are warranted to evaluate the ability of the technology to identify pre-invasive and early-invasive disease."

Rebuttal 1

Reviewer #1 (Report for the authors (Required)):

In this manuscript, the authors reported a nanosensor array composed of multiple ssDNA-encapsulated organic color center-modified carbon nanotubes (OCC-DNAs) together with a machine learning model to achieve perception-based detection of ovarian cancer from patient serum samples. The optimal model in this study yielded 95% specificity, which outperformed the serum biomarker cancer antigen with a specificity of only 81% (CA125). Most importantly, the presented technology required a small sample volume from each patient (360 µL) and relatively short testing time (~ 2 hrs), potentially deriving benefits for clinical screening of ovarian cancers. However, the authors did not explain the underlying mechanisms of the proposed nanosensor array and the specific function of encapsulating ssDNA. Moreover, the sample size was relatively small for an established and reliable prediction model. The authors should discuss more on the biases, limitations, robustness, and generalizability of their model. The following issues also need to be addressed.

We appreciate this meticulous review. These comments prompted multiple new analyses and substantial textual revisions that resulted in a significantly improved manuscript. Below we address all comments and indicate where any associated changes have been made in the revised manuscript.

1. Regarding the design of the study, what is the standard of ssDNA selection for SWNT encapsulation? The authors only mentioned that the ssDNA was chosen based on their ability to form tight structures on the SWCNT surface; however, no explanations were provided regarding their interactions with potential biomarkers of ovarian cancers and the underlying mechanisms behind the selection. Moreover, it seems that the proposed nanosensor was not specific towards ovarian cancer, then how can the proposed achieve accurate screening of ovarian cancer?

Regarding the selection of ssDNA, we appreciate this question and have added more detail on the selection criteria to the introduction of the manuscript and references to several key citations. Our DNA sequence selection is based on the recognition sequences that form specific wrapping patterns on the nanotube surface. These were originally selected to isolate individual (n,m) species/chiralities of nanotubes (Ao *et al J. Am. Chem. Soc.* **2016**). We reasoned that these sequences would confer the greatest diversity of interactions with the serum milieu, which is important to establish an OCC-DNA library for screening disease-specific sensor responses. We based this rationale on the findings that single-stranded DNA encapsulates CNTs via π-π stacking interactions, and certain DNA sequences can behave like a “molecular mask” that defines the shape and size of the exposed surface (Roxbury *et al Nano Lett.* **2012**, Roxbury *et al J. Phys. Chem. B* **2013**). For example, the ss$(TAT)_4$ sequence on the (6,5) nanotube takes the form of single helical wrapping. The sequence ss$(CCG)_2$CC on (6,5) assumes a β-barrel structure. Their characteristic surface structures are responsible for diverse physicochemical properties of the OCC-DNAs, leading to different protein corona compositions (Pinals *et al J. Am. Chem. Soc.* **2020,** Tenzer *et al Nat. Nanotech.* **2013**). The propensity of these resolving sequences to form ordered structures thereby contributes to the different morphology and therefore sensitivity of the nanotube surfaces to the serum milieu. The fluorescence modulation of SWCNTs is caused by several mechanisms including Fermi level shifting through modulation of the immediate redox environment and exciton disruption in response to binding events, which change SWCNT intensity, and solvatochromic (wavelength) shifting due to perturbation of the local dielectric environment, including shifts due to modulation of the local electrostatic environment (Heller *et al PNAS* **2011,** Roxbury *et al ACS Nano* **2016**). Interactions between HGSOC serum biomarkers and OCC-DNA hybrids elicited diverse spectral responses of the sensor array that enabled sufficient differentiation of signals from other sera. We summarized this explanation in the Results and Discussion sections accordingly:

Page 3:
“The ssDNA sequences were chosen based on the recognition sequences of DNA that form specific wrapping patterns on the SWCNT surface (Ao *et al J. Am. Chem. Soc.* **2016**) to result in diverse, highly-defined surface morphologies to confer disparate sensitivities to the local environment (Roxbury *et al Nano Lett.* **2012**, Roxbury *et al J. Phys. Chem. B* **2013**).”

In Page 8:
"We constructed a nanosensor array platform comprised of OCC-DNA elements and coupled with machine learning algorithms, to investigate the potential to identify HGSOC in patient sera. The array was comprised of multiple OCC moieties and DNA sequences, which together offer a rich design space for modulating the morphology and chemistry of the exposed nanotube surface. Our DNA sequence selection was based on the recognition sequences that form specific wrapping patterns on the nanotube surface. These sequences were originally selected to isolate individual (n,m) species/chiralities of nanotubes (Ao *et al J. Am. Chem. Soc.* **2016**). We reasoned that the recognition sequences of DNAs would confer the greatest diversity of interactions with the serum milieu, which is important to establish an OCC-DNA library for screening disease-specific sensor responses. We based this rationale on the findings that ssDNA encapsulates CNTs via π-π stacking interactions, and certain DNA sequences can behave like a "molecular mask" that defines the shape and size of the exposed surface (Roxbury *et al Nano Lett*. **2012**). Their characteristic surface structures are responsible for diverse physicochemical properties of the OCC-DNAs (Roxbury *et al J. Phys. Chem. B* **2013**), leading to different protein corona compositions (Pinals *et al J. Am. Chem. Soc.* **2020**, Tenzer *et al Nat. Nanotech.* **2013**). Different morphologies determined by OCCs and DNA thereby contribute to the selectivity of the nanotube surfaces to the serum milieu. The fluorescence modulation of SWCNTs is caused by several mechanisms including Fermi level shifting through modulation of the immediate redox environment and exciton disruption in response to binding events, which change SWCNT intensity, and solvatochromic (wavelength) shifting due to perturbation of the local dielectric environment, including shifts due to modulation of the local electrostatic environment (Heller *et al PNAS* **2011**, Roxbury *et al ACS Nano* **2016**). OCC fluorescence, on the other hand, is molecularly specific and extremely sensitive to the local chemical environment of the atomic defect sites (Kwon *et al J. Phys. Chem. C* **2015**, Wu *et al Nat. Comnun.* **2019**). Interactions between HGSOC serum biomarkers and OCC-DNA hybrids elicited additional, diverse spectral responses of the sensor array that enabled sufficient differentiation of signals from other sera."

Regarding the question of the specificity of the method to ovarian cancer, we understand the concern. The technology was developed using several moderately-selective sensor elements wherein a sufficient number of elements was included to provide a diverse enough set of responses that could result in the successful training of a machine learning algorithm to recognize the response as either cancer or non-cancer (or at least non-HGSOC). In principle, this method could be used to detect other diseases without necessarily changing the set of nanosensors. However, we did initially select the OCC-DNA nanosensors to be included in the array by narrowing down from a larger library of OCC-DNAs using an initial assessment against a small set of ovarian cancer patient samples (**Fig 1B**). We performed a parametric t-test to identify the statistically significant difference between healthy and HGSOC serum samples in nanosensor responses (quantified by p-value). Our hypothesis was that OCC-DNAs that perform well independently would make good choices when used in combination. Six OCC-DNAs exhibiting E_{11} or E_{11}^- peak wavelengths with statistically significant differences between HGSOC and healthy groups (p-values of < 0.10) were chosen for collective use as an array. The selection/reduction of features improves the training speed and model performance by eliminating redundant features in the data set. We revised the text describing the screening of OCC-DNA elements:

Page 3:
"For each OCC-DNA nanosensor, we analyzed four different spectral features of the OCC-DNA nanosensors that were modulated in response to interactions with analytes in serum: E_{11} and E_{11}^- intensity (*int* and *int**) and wavelength (*wl* and *wl**). From these data, we identified the sensors that gave statistically significant differences in response to normal (healthy control) versus cancer groups in parametric t-tests (quantified by p-value, **Fig 2b** and **Fig S2**). Our hypothesis was that OCC-DNAs which perform well independently would make good choices when used in combination. Six OCC-DNA nanosensors exhibiting E_{11} or E_{11}^- peak wavelengths with statistically significant differences between HGSOC and healthy groups (p-values < 0.10) were selected for the sensor array used in the subsequent parts of this study (highlighted in **Table 1**, **Fig S2**). The selection/reduction of features improves the training speed and model performance by eliminating redundant features in the data set."

2. In some recent studies, machine learning has been deployed to facilitate multi-analyte screening

test combing both ctDNA and protein biomarker tests (Science 2018, 359, 926; Science 2020, 369, eabb9601), which can achieve >99% specificity and >95% sensitivity for ovarian cancer, as well as identification of other types of cancers. Is the presented model cost-effective enough and amenable to high throughput? What are the advantages compared to these methods that appear simpler?

We thank the reviewer for this comment. We have modified the manuscript discussion substantially to discuss the unique features and potential utility of this technology. There are several unique potential benefits that we describe here:

1) This method could be rapidly adapted to the detection of many diseases/conditions. The sensor array could be used to train an algorithm to recognize nearly any condition when given enough data from the sensor responses to the appropriate patient serum samples.
2) This method could supplement or replace the use of known biomarkers when there are issues with selectivity. Due to the potential to iteratively modify the sensor array and machine learning algorithms, and to additively augment training set size, the selectivity may be increasingly optimized.
3) This sensor platform can potentially be used in a high-throughput fashion, due to its relatively low intrinsic costs and minimal need for sample processing, to facilitate the screening of large patient populations.
4) Because the technology does not rely on antibody-based molecular recognition elements, the sensors could be more robust than existing methods and are not as reliant on cold chain logistics, enabling use in resource-limited settings and in technologies such as point-of-care and wearable/implantable devices.

To discuss these potential benefits of the technology, we summarized the discussion above in the main text (pasted below).

Page 9:
"This sensor technology platform exhibits several unique potential advantages for clinical applications. First, this method could be rapidly adapted to the detection of many diseases/conditions. The array could be used to train an algorithm to recognize nearly any disease when given enough data from the sensor responses to the appropriate patient serum samples. Second, this technology could supplement or replace the use of known biomarkers when there are issues with selectivity in conventional multi-analyte tests. Due to the potential to iteratively modify the sensor array and machine learning algorithms and to additively augment training set size, the selectivity may be increasingly optimized. Third, this sensor platform can be used in a high-throughput fashion to facilitate the screening of large populations. Fourth, because the technology does not rely on antibody-based molecular recognition elements, the sensors could be more robust than existing methods and are not as reliant on cold chain logistics, enabling use in resource-limited settings and in technologies such as point-of-care and wearable/implantable devices (Lee *et al ACS Sens.* **2019**)."

3. The sample size of the dataset was 269, including training, validation, and test sets, which was quite small to establish an algorithm with good generalizability in a larger population. Moreover, the authors should also specify how the data were assigned into these partitions and whether there are any systemic differences between the data in each partition.

The reviewer is correct regarding the relatively small number of samples, and we have now changed the manuscript to better address this aspect. Due to the relative rarity of ovarian cancer, the ability to generate training examples for this disease is normally limited to hundreds, even at large cancer centers. To address this relatively sparse data set and evaluate generalizability of the prediction models, we employed cross-validation in algorithm optimization and testing with a new set of samples (collected from different patients and from a later time point) in model validation. In the cross-validation process, the variance of F-score and sensitivity remained small (<0.1), suggesting that the optimized models are generalizable within the sample set. We then collected a separate set of samples, comprising a validation (test) set, which was not used in the training process. With this set, we observed 100% sensitivity, 95% specificity, and an F-score of 0.978 in the optimized SVM model, which values are consistent with the cross-validation results. The results support good generalizability

of the optimized prediction models in a larger population. We clarified our work to rule out overfitting in the revised manuscript; the changes are pasted below.

In Page 5:
"Small variances in F-score and sensitivity (<0.1) in cross-validations suggest that the optimized models are generalizable within the sample set."

"To further assess the robustness of the sensor array and algorithm, we synthesized a new batch of OCC-DNAs under the same condition and collected the sensor array response to an independent test set of 54 patient samples (N_{sa} = 54). To evaluate the model performance in various medical conditions, the test set was sampled from different patients, comprised of 7 HGSOC, 5 other gynecologic diseases, 32 non-gynecologic diseases, and 10 heathy patients. With this new sample set, the optimized SVM model resulted in 100% sensitivity at 98% specificity and an F-score of 0.978. These values are consistent with the cross-validation scores and gave a similar receiver operating characteristic (ROC) curve (**Fig 3d**), indicating that the model did not overfit the data."

Regarding the dataset assignments in these partitions, we conducted a deep risk of bias assessment via Prediction model study Risk Of Bias ASsessment Tool (PROBLAST, Wolff *et al Ann. Intern. Med.* **2019**) to address the reviewer's question of whether there are systematic differences between partitions. Using this tool resulted in the finding of no systemic difference between training and cross-validation sets. In the cross-validation process, stratified shuffle split validation was used to randomly partition the data set into ten subsamples. Each time, nine of the ten subsamples were used to train the model, while a single subsample was used to evaluate the performance of the trained model. The trained models were then tested with a new set of patient samples (test set) to rule out the overfitting issue. The patient samples of the test set were independently collected to validate the prediction models. We note that breast cancers in the non-HGSOC group were enriched in the test set. Table S1 shows patient demographics and biomarker levels to provide a better comparison between the partitions. We listed the risk of bias introduced by the selection of participants and analysis in PROBLAST. The PROBLAST form is now included in Supplemental Information as **Appendix 1** and attached in the end of the rebuttal as well. New discussion has been added to the manuscript and is pasted here:

In Page 5:
"The risk of bias in the study was evaluated based on Prediction model Risk Of Bias Assessment, PROBLAST (Wolff *et al Ann. Intern. Med.* **2019**) (**Appendix 1** in Supplemental Information). The risk of bias scored low in terms of predictors, outcomes, and analysis. In participants, the tool resulted in the finding of no systemic differences between training and cross-validation sets."

Serum sample set in Methods (Page 10)
"Patient demographics, diagnosis, and biomarker levels are available in Table S1."

Model training and performance assessment in Methods (Page 11)
"In the cross-validation process, stratified shuffle split validation was used to randomly partition the data set into ten subsamples. In each partition, nine of the ten subsamples were used to train the model, while a single subsample was used to test the trained model. The average F-score of the ten-fold cross-validation was used to assess model performance. The trained models were then tested with an independent set of patient sera (N=54), sampled from different patients (test set), as external validation."

4. The authors classified the patient population into three categories – HGSOC, other non-HGSOC diseases, and healthy individuals. From the biological point of view, there may be a wide discrepancy in the test results among the non-HGSOC disease category, especially when low-grade ovarian carcinoma and some other cancers are included. The generalization of the presented model is doubtful when applying to real scenario screening.

This is a keen and helpful comment which prompted us to revisit the tri-categorial classification model. As the non-HGSOC group covered a variety of disease types, the classification accuracy of the non-

HGSOC category can significantly vary depending on the training set. Classifying an “other disease” group in our data set does not have clinical utility nor does it represent any characteristic sensor response that covers the wide variation of the disease types. Particularly, our tri-categorial classification model of “other diseases” used a substantial number of breast and endometrial cancer samples due to their abundance in the patient population that we measured, which one could imagine would bias a rule-out decision based on these diseases. Thus, we decided to use the combined group comprising “normal” and “other diseases” and remove the tri-categorical classification. We thus removed **Fig 3g** and the corresponding discussion from the manuscript.

5. Although the authors pointed out the potential of the technology for clinical translation and biomarker discovery, the authors did not mention 1) how the proposed technology can improve the performance and efficiency of clinicians; 2) how the feature values influence the prediction outcomes; 3) and how these features can link to the underlying biology of ovarian cancer.

Regarding point #1: We agree that this should be better discussed, and we have modified the manuscript to elaborate on how the sensor technology can benefit the efficiency of clinicians. This technology has the potential to be developed as a screening tool to result in a single, easy-to-interpret test result in primary care settings. A positive result would then prompt referral for follow-up care. Current clinical practice includes longitudinal CA125 measurement and second-line transvaginal ultrasonography. Compared to this multimodal screening, the technology that we described enables a rapid and inexpensive screening with high clinical sensitivity (87% sensitivity at 98% specificity), ultimately improving the ability of clinicians to make a rapid and accurate diagnosis of a disease for which lead time is paramount. As discussed in response to Comment #2 above, the technology could be used in resource-limited settings, potentially as a point-of-care device, enabling screening in large populations. As a potential later development, the technology could be incorporated into/compatible with wearable technologies, although one would not expect that different tests, instead of HGSOC, would be used in this form factor. We have modified the text to address this issue in the revised manuscript and pasted that text below:

In Page 8:
“Model performance of the sensor technology exceeds the results of current best clinical screening test using longitudinal CA125 and second-line transvaginal ultrasonography (Jacobs *et al Lancet* **2015**) (87% vs. 84% clinical sensitivities at 98% specificity).”

Page 9:
“Fourth, because the technology does not rely on antibody-based molecular recognition elements, the sensors could be more robust than existing methods and are not as reliant on cold chain logistics, enabling use in resource-limited settings and in technologies such as point-of-care and wearable/implantable devices (Lee *et al ACS Sens.* **2019**). Lastly, the sensor technology also has the potential to be developed as an inexpensive and rapid screening tool to result in a single, easy-to-interpret test result in primary care settings.”

Regarding point #2: We thank the reviewer for bringing up the important question. To address how the feature values influence the prediction outcomes, we conducted a new set of analyses of the relative importance of each feature value using an ablation study. We added this to the manuscript and also included discussion about how the inclusion or removal of certain features change the prediction outcomes. The feature importance study examined how the prediction scores change when a single feature was dropped from the data set. We found that the same feature in different sensor arrays can improve or reduce the prediction scores. For example, the E_{11}^- intensity (*dint**) of NEt_2*CTT exhibited the highest positive feature importance (0.067) in the sensor array of 4-$N(C_2H_5)_2$*$CT_2C_3T_2C$, indicating the inclusion of *dint** in the machine learning model increased the F-score from 0.8359 to 0.9029. On the other hand, the same feature reduced the model performance by 0.018 in the sensor array combination of 4-$N(C_2H_5)_2$*$CT_2C_3T_2C$ and 3,4,5-F_3*$(GT)_{15}$. However, we note that it is still unclear exactly which physicochemical interactions cause the spectral responses of each sensor element and how the spectral parameters can be systematically designed to improve the sensor array combination. To report the relationship between feature values and model performance, we have

added the feature importance analysis of each spectral variable in all OCC-DNA sensors in the arrays used for this work as a panel in **Fig S9** of the revised manuscript (and pasted below), and we modified the main text to explain accordingly.

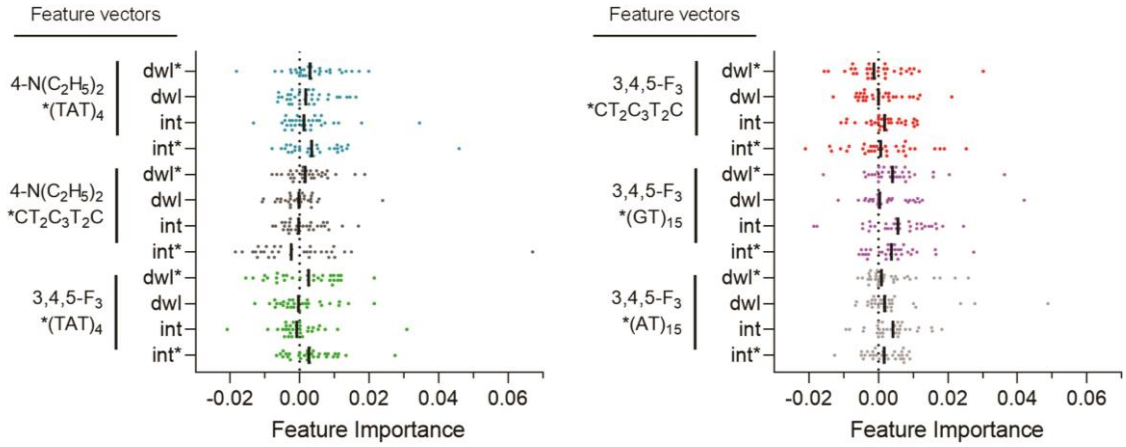

Fig S9 | Relative feature importance of each spectroscopic variable in the HGSOC binary classification models. a, Feature importance of each spectral parameter, used to train the SVM models, of all OCC-DNA sensors in the arrays tested in this work. Solid lines indicate the median feature importance.

Page 9:
"We also found that the same feature in different sensor arrays can improve or reduce the prediction scores (**Fig S9**). For instance, the E_{11}^- intensity (*dint**) of NEt_2*CTT has the highest positive feature importance (improved F-score by 0.067) in the sensor array of 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ while the same feature reduced the F-score by 0.018 in the sensor array combination of 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ and 3,4,5-F_3*$(GT)_{15}$."

Regarding the link between the features and the underlying biology of ovarian cancer (point #3), we believe that this technology may serve as a tool to uncover new information on the pathophysiology of ovarian cancer, starting with the identification of markers that are modulated upon disease initiation and progression. Although machine learning itself currently results in predictions via a fairly opaque process (although investigators are attempting to rectify this) (Zednik C *Philos Technol* **2021**), protein corona analysis may inform which serum proteins constitute the fingerprint and are responsible for disease-specific sensor responses. Proteome analysis can be employed to study the protein corona composition on each sensor element. To address how the sensor technology can potentially benefit the biology of ovarian cancer, we modified the manuscript text as follows:

Page 9:
"It is important to note that the best-performing HGSOC prediction model (**Fig 3d**) included the spectroscopic variables that were not sensitive to the known biomarkers and their relative importance was much significant than the biomarker-related variables (**Fig S9**). This suggests that there exist potential biomarkers or combinations thereof that are either unknown or not part of conventional screening approaches but were captured by the OCC-DNA sensor array. Information detailing which biomarkers and molecular interactions primarily result in the disease fingerprint is unknown and largely cannot be determined by current machine learning methods (Zednik C *Philos Technol* **2021**). We believe that it may be possible, with extensive investigations, to use quantitative proteomics aided by the nanosensor array as a discovery tool (Docter *et al Nat Protoc* **2014**, Lai *et al ACS Nano* **2012**, Pinals *et al Angew Chem* **2020**). Such investigation could potentially be used to facilitate biomarker discovery efforts (Hadjidemetriou, M. et al *Nano Today* **2020**) and uncover new information related to disease pathophysiology."

6. The authors should compare the performance between the presented model to previously

published models to show its superiority. Moreover, apart from the comparison of specificity between the presented model with the CA125 test, the estimated screening cost should also be compared.

We thank the reviewer for the suggestion to compare our sensor technology with other models. Despite many promising new markers for ovarian cancer, CA125 remains the single-best biomarker (Cramer *et al Cancer Prev Res* **2011**, Moore *et al Cancer* **2012**). We thereof compared the sensitivity at 98% specificity between our nanosensor array and previously reported models with the largest clinical screening trial using the combination of longitudinal CA125 and TVS (Menon *et al Lancet* **2021**). Our optimized model exhibited >87% sensitivity at 98% specificity for HGSOC prediction that is comparable with the largest multi-cohort screening efforts (84% clinical sensitivity). We revised the result section to incorporate the comparison.

Page 8:
"Model performance of the sensor technology exceeds the results of current best clinical screening test using longitudinal CA125 and second-line transvaginal ultrasonography (Jacobs *et al Lancet* **2015**) (87% vs. 84% clinical sensitivities at 98% specificity)."

Regarding the screening cost, we estimated that our sensor test would cost much less than the multiplexing of conventional markers judging from a CA125 blood test ($50 or higher per single-analyte test, if the screening accompanied with TVS and/or multiple biomarkers). The low screening cost is also attributed to: 1) the lack of a need for cold chain storage required due to the high physical stability of the sensors, and 2) the small amount of OCC-DNAs needed for screening (<5 nanograms). Currently, the materials needed for the sensor average approximately $5 per sample. The cost of the sensor measurement would also diminish if measured via high-throughput devices, and the technology can be built to use very low sample volumes. We have added this new discussion in the manuscript.

Page 9:
"Lastly, the sensor technology also has the potential to be developed as an inexpensive and rapid screening tool to result in a single, easy-to-interpret test result in primary care settings. The materials needed for the sensor cost approximately $5 per sample due to the lack of need for cold chain storage and the small amount of OCC-DNAs needed for screening (<5 nanograms). The cost of the sensor measurement would also diminish if measured via high-throughput instruments, and the potential for the use of very low sample volumes is substantial."

7. The risk of bias should be assessed in terms of participants, predictors, outcomes, and analysis.

We appreciate the reviewer's suggestion to evaluate the risk of bias of our study. To address this issue, we assessed potential shortcomings in study design, conduct, or analysis using the Prediction model Risk Of Bias ASsessment Tool (PROBLAST, Wolff *et al Ann. Intern. Med.* **2019**). Based on PROBLAST, no significant risk of bias was identified in all domains except for participants. We identified high risk of bias in terms of participants – inclusion and exclusion of participants may not be appropriate in our study. The rationale for the high risk of bias is the limited patient information for healthy donors, including medication, hysterectomy, and age distribution, and enriched fraction of breast cancers in the non-HGSOC group of the test set. In terms of predictors and outcomes, the risk of bias scored low because they appear to match all the review questions. In analysis, our study was judged to be at low risk of bias as well. Comorbidities affect the blood constituents and such chemical and biological interferents could interact with the nanosensors, potentially introducing interferences in HGSOC-specific sensor response. However, the analyses suggested no indications that such interferences reduced specificity in HGSOC detection by the sensor platform (see the response to Comment #9). We think that the assessment will help the readers better understand our experiment design and provide the valuable information about which factors need to be considered for the clinical translation of this technology in the future. We have added the discussion of sources of potential bias and how we addressed the risks of bias in the manuscript. The PROBLAST form has been added to the Supplemental Information as **Appendix 1** (and is pasted at the end of the rebuttal).

In Page 5:

"The risk of bias in the study was evaluated based on Prediction model Risk Of Bias Assessment, PROBLAST (Wolff *et al Ann. Intern. Med.* **2019**) (**Appendix 1** in Supplemental Information). The risk of bias scored low in terms of predictors, outcomes, and analysis. In participants, the tool resulted in the finding of no systemic differences between training and cross-validation sets. However, the limited medical record of healthy donors and enriched fraction of breast cancers in the non-HGSOC group of the test set may introduce systematic bias in participant selection and the validation of machine learning models, respectively. For clinical translation of the technology, these risk of bias must be taken into account."

8. According to Figure 2b and Figure S2, the spectral differences of E11 and E11' between HGSOC and benign samples were only within 2 nm in many groups. Such a slight difference was not sufficient to distinguish the pathological status because many factors (e.g., sample preparation, ion concentration, etc.) can cause small errors in spectral measurements.

We understand the reviewer's concern that wavelength differences of ~ 2nm of OCC-DNAs appear small. We conducted additional data analysis to address this issue. To reduce the potential inconsistency in spectral measurements that may result from sample processing, we measured each patient sample in triplicate, and the averaged sensor response was used for the data analysis as noted in Methods: Data preprocessing. We analyzed the variation of each measurement from the averaged triplicates (see the pasted histograms below). For all the OCC-DNAs peaks, *dwl* and *dwl** showed narrow Gaussian distributions and the standard deviations ranged from 3.72–5.37%. The maximum variation in the same sample was less than 15% (<0.3 nm). The analysis confirmed that our measurement and data preprocessing can reliably identify the small spectral shifts. Furthermore, we would like to note that the combination of spectral parameters elicited significant segregation between HGSOC and benign groups in principal component analysis (**Fig 2d**). We note that these differences may be significantly better than expected, especially as compared to organic fluorophore, because OCC-DNAs exhibit the narrow full-width-at-half-maximum (FWHM) of the SWCNT fluorescence bands (35–80 meV), and thus, 2 nm shifts are significantly easier to be resolved with SWCNTs than conventional fluorophores (>100 meV). Below, we provide the frequency distributions of variation in *dwl* and *dwl** for each measurement of OCC-DNA emission. We have summarized these analyses and the figure below is included as **Fig S3**. A new discussion of this issue is now provided in the text and pasted below.

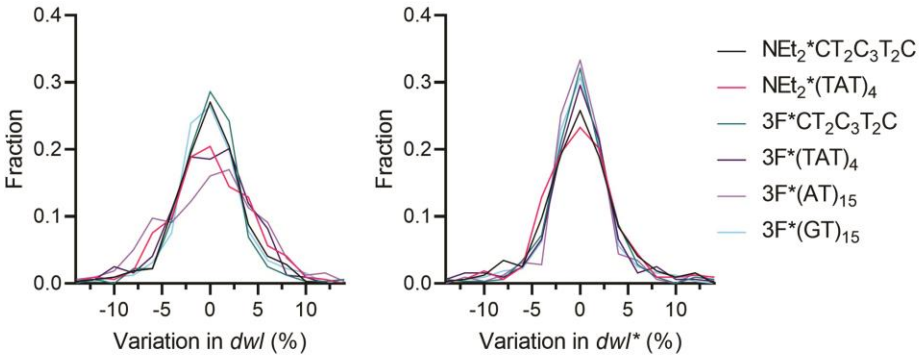

Fig S3. Frequency distribution of variation in E_{11} and E_{11}^- wavelength shifts. The variation was calculated as the difference between the center wavelength of each measurement from the average of the triplicate measurements.

In Page 4:
"To reduce the inconsistency in spectral measurements, the averaged sensor response of triplicate was used for the data analysis. We note that the variation of each measurement from the averaged triplicates was small for all the OCC-DNA peaks (**Fig S3**). The variations in *dwl* and *dwl** showed narrow Gaussian distributions with the standard deviations ranged from 3.72–5.37%. The maximum variation in the same sample was less than 15% (<0.3 nm). The analysis confirmed that our measurement can reliably identify the small spectral shifts. This is likely because OCC-DNAs exhibit

relatively narrow bandwidths (35–80 meV), and thus, small spectral shifts are significantly easier to be resolved as compared to conventional fluorophores (>100 meV).”

9. Following the previous question, the specificity of such sensor to detect cancer under the backdrop of other diseases is doubted. For instance, to detect ovarian cancer in aged patient or patient with autoimmune diseases, such as diabetes? These background chronic diseases can affect the blood constituents. The authors need to validate this.

The reviewer raised an important point. Background chronic diseases could affect the blood constituents, introducing interferences in sensor response. To address this concern, we conducted a new thorough chart analysis to identify comorbidity based on the patients’ medication information and checked the correlation with the model performance. From the chart review, 12 different chronic conditions were identified in the studied patients including allergies, depression, high blood pressure, and diabetes. 75% and 68% of HGSOC and other disease patients suffered from at least one chronic condition. The relative abundance between these disease groups was similar (see the figure pasted below, and added into the manuscript as **Fig S6b**). We confirmed that our prediction model can accurately classify HGSOC regardless of patients’ chronic conditions. Based on these analyses, we did not find strong indications for reduced specificity in HGSOC detection in our sensor platform.

The medication information also relates variance in blood constituents by background chronic diseases. Although all the medications listed do not accumulate in body and the concentration in blood is low, they potentially interfere with the spectral response of OCC-DNA nanosensors. We thus identified small-molecule drugs whose occurrence differs by 0.1 or higher between HGSOC and other disease groups, and statistically assessed the contribution of each interferent to the sensor response by multivariate regression (**Table S5**). The adjusted R^2 of the regression model ranged from -0.045 to 0.233, indicating weak correlations of each sensor response to medications. We also confirmed that the sensor platform accurately classified the disease status regardless of medication, evidenced by high F-scores of the HGSOC screening models, suggesting minimal interference to the disease prediction.

Lastly, with enough patients in the training set, all with different comorbidities would converge as background, the disease-specific signal rises above the noise. The algorithm will learn from and adjust for patients who are taking different drugs and have different comorbidities.

We have incorporated this discussion into the main text and PROBLAST analysis. We also included the distribution of chronic disease conditions figure as panels in **Fig S6**.

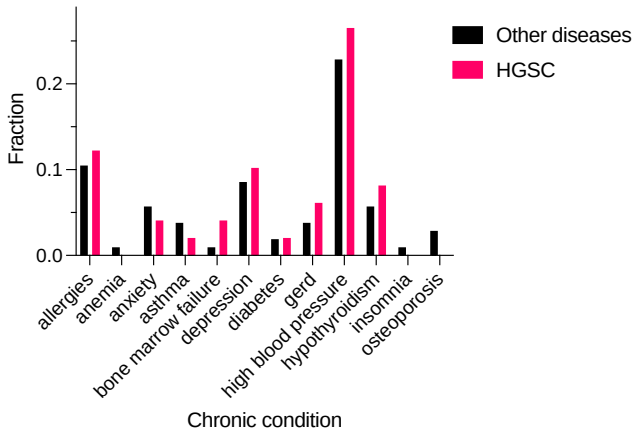

Fig S6 | Assessment of medications as potential interferents to nanosensor prediction. b, Chronic conditions, and prevalence thereof, in patients measured in this study. Comorbidity was identified based on the patients’ medication information.

"We also endeavored to account for chemical interferents and background chronic conditions that could confer a bias in the sensor response. From a patient chart review, we identified chronic diseases and most common medications administered to the patients (**Fig S6**). We found that 75% of HGSOC and 68% of other disease patients suffered from at least one chronic condition, and the relative abundance between these disease groups was similar. Regarding the medications, we statistically assessed the contribution of each interferent to the sensor results using a multivariate regression model (**Table S5**). The regression model determined a linear correlation between the sensor response and medication, using estimated parameters and errors. The adjusted R^2 of the regression model ranged from -0.045 to 0.233, indicating weak linear correlations of each sensor response to medications. We confirmed that the sensor array platform accurately classified the disease status regardless of medication and chronic conditions, evidenced by high F-scores of the HGSOC prediction models (**Table S3**). The analysis suggested no indications that such interferences reduced specificity in HGSOC detection by the sensor platform."

10. The authors only validated the performance using the test data from the same source. It is better to include both internal and external validation to provide more insight into the generalization of the model.

We agree that external validation de-risks any potential systematic distortion in the prediction model estimates. We included both internal and external validations to develop a robust prediction model based on PROBLAST (Wolff *et al Ann. Intern. Med.* **2019**). To minimize the generalization error, we employed cross-validation in algorithm optimization as internal validation. This is internal because the same data are used for both development and validation. We then quantified the model performance in data external to the development samples. We tested a new sample set that was acquired from new patients at a later timepoint as external validation. In the cross-validation process, the variance of F-score and sensitivity remained small (<0.1), suggesting that the optimized models are generalizable within the sample set. With the new sample set, we observed 100% sensitivity at 98% specificity and an F-score of 0.978 in the optimized SVM model, which values are consistent with the internal validation results. External geographic validation would require other institutions/locations to use the sensors and specialized instrumentation that we developed, which is not yet possible. We developed a unique array of OCC-DNA sensors. OCC-DNA complexes have not yet been synthesized or published anywhere else to date. We also developed a specialized instrument for high-throughput near-infrared spectroscopy with 730 nm excitation, which is not yet available in other laboratories. We have now discussed these issues in the revised manuscript.

In Page 5,
"Small variances in F-score and sensitivity (<0.1) in cross-validations suggest that the optimized models are generalizable within the sample set."

"To evaluate the model performance in various medical conditions, the test set was sampled from different patients, comprised of 7 HGSOC, 5 other gynecologic diseases, 32 non-gynecologic diseases, and 10 heathy patients. With this new sample set, the optimized SVM model resulted in 100% sensitivity at 98% specificity and an F-score of 0.978. These values are consistent with the cross-validation scores and gave a similar receiver operating characteristic (ROC) curve (**Fig 3d**), indicating that the model did not overfit the data."

11. In Figure S3, "a", "b", "c" and "d" appeared in the figure caption; however, they can no be found the Figure. The authors are recommended to check the figures carefully to make sure there is no missing or errors in each figure.

We appreciate the reviewer for pointing out this oversight. We double-checked the figure captions and figures and revised them accordingly.

12. The authors claimed that there is linear correlation of each sensor response to medications in Figure S5. However, we did not find any linear correlation in Figure S5. How did the authors make this

conclusion?

We thank the reviewer for pointing this issue out. Figure S5 presents the occurrence of medications between HGSOC and other disease groups. The statistical assessment of the contribution of interferents to the sensor responses is associated with **Table S5**. We employed a multivariate regression model to identify the correlation between medication and sensor response. The low R^2 values in the regression model indicate no significant correlation, suggesting that medications did not significantly impact the sensor response. To avoid confusion, we modified the discussion and reference to **Table S5**:

In Page 6,
"We also endeavored to account for chemical interferents and background chronic conditions that could confer a bias in the sensor response. From a patient chart review, we identified chronic diseases and most-common medications administered to the patients (**Fig S6**). We found that 75% of HGSOC and 68% of other disease patients suffered from at least one chronic condition and the relative abundance between these disease groups was similar. Regarding the medications, we statistically assessed the contribution of each interferent to the sensor results using a multivariate regression model (**Table S5**). The regression model determined a linear correlation between the sensor response and medication, using estimated parameters and errors. The adjusted R^2 of the regression model ranged from -0.045 to 0.233, indicating weak linear correlations of each sensor response to medications. We confirmed that the sensor array platform accurately classified the disease status regardless of medication and chronic conditions, evidenced by high F-scores of the HGSOC prediction models (**Table S3**). The analysis suggested no indications that such interferences reduced specificity in HGSOC detection by the sensor platform."

Reviewer #2 (Report for the authors (Required)):

The authors presented a well written article for "Machine Perception Nanosensor Array Captures a Disease Fingerprint of Ovarian Cancer". They report a machine model of a nanosensor array to collect physiochemical interactions to differentiate ovarian cancer from other diseases as well as healthy individuals. The report of sensitivity and selectivity are impressive. However, the reviewer has concerns about the nanosensor ability to recognize and differentiate high-grade serous ovarian carcinoma (HGSOC).

We appreciate the thorough review of our manuscript and the reviewer's recommendations for improvement. These critiques prompted multiple edits that substantially improved our manuscript. Below we address this critique and indicate where any associated changes have been made in the revised submission.

1. Could the authors elaborate on how and why these specific DNA sequences were used. As noted in the narrative, it is because they interact with the single walled carbon nanotubes (SWCNT). However, what is causing the sensor to recognize biological elements in the patient sample? It is not clear to the reader specifically how these sequences are used to differentiate samples particularly for HGSOC vs non-HGSOC. For example, an antibody on a sensor surface can selectively bind to a protein. Do these DNA sequence act as aptamers? What is leading to this high selectivity?

Regarding the selection of ssDNA, we appreciate this question and have elaborated the selection criteria to the introduction of the manuscript and references to several key citations. Our DNA sequence selection is based on the recognition sequences that form specific wrapping patterns on the nanotube surface. These were originally selected to isolate individual (n,m) species/chiralities of nanotubes (Ao *et al J. Am. Chem. Soc.* **2016**). We reasoned that these sequences would confer the greatest diversity of interactions with the serum milieu, which is important to establish an OCC-DNA library for screening disease-specific sensor responses. We based this rationale on the findings that single-stranded DNA encapsulates CNTs via π-π stacking interactions, and certain DNA sequences

can behave like a “molecular mask” that defines the shape and size of the exposed surface (Roxbury *et al Nano Lett.* **2012**, Roxbury *et al J. Phys. Chem. B* **2013**). For example, the ss$(TAT)_4$ sequence on the (6,5) nanotube takes the form of single helical wrapping. The sequence ss$(CCG)_2$CC on (6,5) assumes a β-barrel structure. Their characteristic surface structures are responsible for diverse physicochemical properties of the OCC-DNAs, leading to different protein corona compositions (Pinals *et al J. Am. Chem. Soc.* **2020,** Tenzer *et al Nat. Nanotech.* **2013**). The propensity of these resolving sequences to form ordered structures thereby contributes to the different morphology and therefore sensitivity of the nanotube surfaces to the serum milieu. The fluorescence modulation of SWCNTs is caused by several mechanisms including Fermi level shifting through modulation of the immediate redox environment and exciton disruption in response to binding events, which change SWCNT intensity, and solvatochromic (wavelength) shifting due to perturbation of the local dielectric environment, including shifts due to modulation of the local electrostatic environment (Heller *et al PNAS* **2011,** Roxbury *et al ACS Nano* **2016**). Interactions between HGSOC serum biomarkers and OCC-DNA hybrids elicited diverse spectral responses of the sensor array that enabled sufficient differentiation of signals from other sera. We summarized this explanation in the Results and Discussion sections accordingly:

Page 3:
“The ssDNA sequences were chosen based on the recognition sequences of DNA that form specific wrapping patterns on the SWCNT surface (Ao *et al J. Am. Chem. Soc.* **2016**) to result in diverse, highly-defined surface morphologies to confer disparate sensitivities to the local environment (Roxbury *et al Nano Lett.* **2012**, Roxbury *et al J. Phys. Chem. B* **2013**).”

In Page 8:
“We constructed a nanosensor array platform comprised of OCC-DNA elements and coupled with machine learning algorithms, to investigate the potential to identify HGSOC in patient sera. The array was comprised of multiple OCC moieties and DNA sequences, which together offer a rich design space for modulating the morphology and chemistry of the exposed nanotube surface. Our DNA sequence selection was based on the recognition sequences that form specific wrapping patterns on the nanotube surface. These sequences were originally selected to isolate individual (n,m) species/chiralities of nanotubes (Ao *et al J. Am. Chem. Soc.* **2016**). We reasoned that the recognition sequences of DNAs would confer the greatest diversity of interactions with the serum milieu, which is important to establish an OCC-DNA library for screening disease-specific sensor responses. We based this rationale on the findings that ssDNA encapsulates CNTs via π-π stacking interactions, and certain DNA sequences can behave like a “molecular mask” that defines the shape and size of the exposed surface (Roxbury *et al Nano Lett*. **2012**). Their characteristic surface structures are responsible for diverse physicochemical properties of the OCC-DNAs (Roxbury *et al J. Phys. Chem. B* **2013**), leading to different protein corona compositions (Pinals *et al J. Am. Chem. Soc.* **2020**, Tenzer *et al Nat. Nanotech.* **2013**). Different morphologies determined by OCCs and DNA thereby contribute to the selectivity of the nanotube surfaces to the serum milieu. The fluorescence modulation of SWCNTs is caused by several mechanisms, including Fermi level shifting through modulation of the immediate redox environment and exciton disruption in response to binding events, which change SWCNT intensity, and solvatochromic (wavelength) shifting due to perturbation of the local dielectric environment, including shifts due to modulation of the local electrostatic environment (Heller *et al PNAS* **2011**, Roxbury *et al ACS Nano* **2016**). OCC fluorescence, on the other hand, is molecularly specific and extremely sensitive to the local chemical environment of the atomic defect sites (Kwon *et al J. Phys. Chem. C* **2015**, Wu *et al Nat. Comnun.* **2019**). Interactions between HGSOC serum biomarkers and OCC-DNA hybrids elicited additional, diverse spectral responses of the sensor array that enabled sufficient differentiation of signals from other sera.”

Regarding the question of the sensitivity of OCC-DNAs to identify ovarian cancer, we understand the concern. The technology was developed using several moderately-selective sensor elements wherein a sufficient number of elements was included to provide a diverse enough set of responses that could result in the successful training of a machine learning algorithm to recognize the response as either cancer or non-cancer (or at least non-HGSOC). In principle, this method could be used to detect other diseases without necessarily changing the set of nanosensors. However, we did initially select the OCC-DNA nanosensors to be included in the array by narrowing down from a larger library of OCC-DNAs using an initial assessment against a small set of ovarian cancer patient samples (**Fig 1B**). We performed a parametric t-test to identify the statistically significant difference between healthy and

HGSOC serum samples in nanosensor responses (quantified by p-value). Our hypothesis was that OCC-DNAs that perform well independently would make good choices when used in combination. Six OCC-DNAs exhibiting E_{11} or E_{11}^- peak wavelengths with statistically significant differences between HGSOC and healthy groups (p-values of < 0.10) were chosen for collective use as an array. The selection/reduction of features improves the training speed and model performance by eliminating redundant features in the data set. We revised the text describing the screening of OCC-DNA elements:

Page 3
"For each OCC-DNA nanosensor, we analyzed four different spectral features of the OCC-DNA nanosensors that were modulated in response to interactions with analytes in serum: E_{11} and E_{11}^- intensity (*int* and *int**) and wavelength (*wl* and *wl**). From these data, we identified the sensors that gave statistically significant differences in response to normal (healthy control) versus cancer groups in parametric t-tests (quantified by p-value, **Fig 2b** and **Fig S2**). Our hypothesis was that OCC-DNAs perform well independently would make good choices when used in combination. Six OCC-DNA nanosensors exhibiting E_{11} or E_{11}^- peak wavelengths with statistically significant differences between HGSOC and healthy groups (p-values < 0.10) were selected for the sensor array used in the subsequent parts of this study (highlighted in **Table 1**, **Fig S2**). The selection/reduction of features improves the training speed and model performance by eliminating redundant features in the data set"

2. It is stated that "to further assess the robustness of the sensor array and algorithm, we synthesized a new batch of OCC-DNAs.", not clear here when you synthesized a new batch was this under the same conditions?

We thank the reviewer for pointing out the lack of clarity in synthesizing OCC-DNAs for validation of this nanosensor array. The new batch of OCC-DNAs was synthesized under the same conditions as the first batch but several months after the first. The interrogation of the sensor array with the newly collected patient serum samples was then conducted 12 months after the collection and processing of the training set data. We clarified the sentence as follows:

Page 5
"To further assess the robustness of the sensor array and algorithm, we collected the sensor array response to a test (validation) set of 54 new patient samples (N_{sa} = 54), and synthesized a new batch of OCC-DNAs under the same conditions"

3. The authors suggest that unidentified biomarkers in serum are responsible for the sensor response as machine model detection of current ovarian serum markers resulted in high F-scores. This seems very interesting and also puzzling. Could the authors elaborate on their model to suggest why biomarker-dependent features scored highly? What are you detecting and how can it be validated through more traditional methods?

We are glad to elaborate on the correlation between biomarker-dependent features and model performance, and we modified the manuscript to clarify this point. We found that several OCC-DNA spectral responses quantitatively correlated with certain known biomarker concentrations within the biologically relevant range upon interrogating them directly (**Fig 4**). The observations indicate that 1) OCC-DNA fluorescence transduces broad types of subtle differences in physicochemical properties of physisorbed molecules, and *2)* known serum biomarkers contribute to the responses of the sensor array that comprise the disease fingerprint. This result explains why the data set including biomarker-dependent features scored highly in the feature importance analysis of HGSOC prediction models and how the machine learning models can be trained to detect an abnormal level of disease biomarkers in serum with the data set. All of this data was benchmarked against conventional immunoassay (**Fig S7**). We clarified the main text to discuss the relation between the high model performance and biomarker-dependent features. The caption of **Fig S7** is now updated to clarify how the data were taken.

Page 7
"Also, the biomarker-dependent features scored highly, indicating that such features improved the SVM model performance (**Fig 4e**). The observations confirmed that OCC-DNA fluorescence transduces broad types of subtle differences in physicochemical properties of physisorbed molecules and *2)* known serum biomarkers make up part of the disease fingerprint."

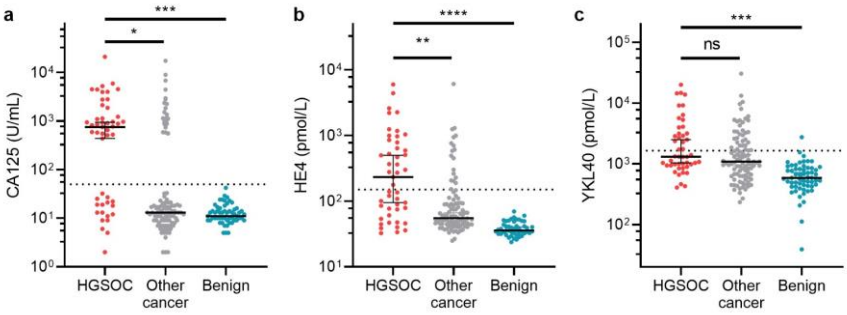

Fig. S1 | Serum levels of known ovarian cancer biomarkers in the model study population. a, CA125, **b,** HE4, and **c,** YKL40. The serum protein levels were quantified by automated immunoassay. Dotted lines indicate the clinical reference of each biomarker for HGSOC diagnosis. The error bars denote median ± 95% CI.

To examine the analytes the sensors detect, we think that quantitative proteome analysis is required. Proteome analysis can quantify protein coronas on OCC-DNA nanosensors and determine which proteins are most responsible for HGSOC identification via the nanosensor array response. We now explained how the protein biomarkers that induce disease-specific spectral responses of the sensor array can be quantified.

Page 8:
"Information detailing which biomarkers and molecular interactions primarily result in the disease fingerprint are unknown and largely cannot be determined by current machine learning methods (Zednik C *Philos Technol* **2021**). We believe that it may be possible, with extensive investigations, to use quantitative proteomics aided by the nanosensor array as a discovery tool (Docter *et al Nat Protoc* **2014**, Lai *et al ACS Nano* **2012**, Pinals *et al Angew Chem* **2020**). Such investigation could potentially be used to facilitate biomarker discovery efforts (Hadjidemetriou, M. et al *Nano Today* **2020**) and uncover new information related to disease pathophysiology."

Interestingly, the biomarker-dependent features did not account for all of the prediction scores. The inclusion of certain features that exhibited no quantitative correlation with known biomarkers improved the model performance. This suggests that the nanosensor array results may be due to the transduction of unidentified biomarkers. Detection of the unknown biomarkers that elicit disease-specific sensor responses can be investigated by quantitative proteomic analysis as explained above. We revised the text to clarify the observations.

Page 7:
"However, the use of biomarker-dependent features exclusively did not result in the best F-scores. The inclusion of certain features that showed no quantitative correlation with known biomarkers improved the model performance. These experiments suggest that the OCC-DNA nanosensor array results may be due, at least in part, to the transduction of heretofore unidentified biomarkers."

4. It is noted in support vector regression models that there was a prediction error in the normal concentration range for CA125 was larger than the high concentration range and it was attributed to the detection limit in single titration experiments. Can the authors elaborate on the detection limits that were selected in their model and how it relates to the clinically significant ranges for the known biomarkers?

This comment prompts us to clarify our methodology regarding our development of support vector regression (SVR) models to quantify the biomarkers such as CA125. The SVR models of serum protein levels were optimized such that the variation of the predicted values from the actual protein levels is minimized (high R-squared). Although there was a linear correlation between the predicted and actual protein levels (R-squared of 0.719 for the best performing SVR model, **Fig 4i**), the optimized SVR models showed some variation in the prediction errors. Due to the variation in prediction errors, we cannot measure the detection limit of SVR.

Regarding the relation to the clinically significant ranges of the known biomarkers, the SVR models did not reliably identify the high levels of biomarkers above the clinical threshold of suspicion of malignancy. Especially for the samples with low biomarker concentrations (CA125 <50 U/mL), the prediction errors can be as high as a factor of 2. Based on the analyses, we concluded that the SVR models suggest that the known biomarkers constitute the sensor responses, but it was not sensitive enough to reliably predict the exact biomarker levels. We modified the manuscript to clarify the prediction accuracy of the SVR models and its implications.

Page 7:
"We additionally investigated whether support vector regression (SVR) models can quantitatively predict serum biomarker levels using the sensor array (**Fig 4h**). The best CA125 regression model, using three OCC-DNAs, NEt_2*$(TAT)_4$, 3F*$(TAT)_4$, and 3F*$(AT)_{15}$ resulted in an average R-squared (r^2) value of 0.719 (**Fig 4i**). We note that the prediction error in the normal concentration range (<50 U/mL) was larger than in the high concentration range. This can be attributed to the fact that the detection limit in the single titration experiment was close to the clinical reference of CA125. The SVR models of HE4 and YKL40 were also constructed, resulting in r^2 values of 0.55 and 0.56, respectively. The SVR models suggest that the known biomarkers constitute the sensor responses, but the models were not sensitive enough to reliably predict the exact biomarker levels."

5. Is this sensitivity and selectivity only from the perception-based model? If so, can the authors elaborate on the specifics of algorithm and what is unique about it? Also, the reviewer questions such a low sample size (269 patients) with this shotgun approach.

We are happy to address the reviewer's questions on our perception-based sensing. As we answered in Comment #1, the high sensitivity and selectivity derive from the use, in concert, of sensors that as individuals are only moderately selective. Our approach is based neither on highly specialized algorithms – we used standard machine learning methods – nor on highly selective individual OCC-DNA sensor elements. For robustness, we investigated a number of machine learning models with nested levels of optimization processes: model hyperparameters, model choice, and multilevel validation. We have shown that a set of optimization processes can classify HGSOC accurately. We now clarified how we optimized the machine learning models for perception-based sensing in the main text.

In Page 4:
"To differentiate HGSOC from other conditions, we next trained machine learning models using the sensor responses and clinical diagnostic results (**Fig 3**, **Fig S5**). The algorithms were used for binary classification of sensor responses: HGSOC vs. other diseases + healthy (the differentiation of HGSOC from all other samples). The set of features chosen for the classification task were the spectroscopic variables *dint*, *dint**, *dwl*, and *dwl** collected from the OCC-DNA sensor array. For robustness, we investigated five standard machine learning algorithms with nested levels of optimization processes: model hyperparameters, model choice, and multilevel validation. We tested supervised machine learning algorithms: logistic regression, decision tree, artificial neural networks, random forest, and support vector machine (SVM) while tuning models' hyperparameters with Bayesian optimization. The averaged F-score in 10-fold cross-validation was used to assess the model performance (see Methods)."

The reviewer is correct regarding the relatively small number of samples, and we have now changed the manuscript to better address this aspect. Due to the relative rarity of ovarian cancer, the ability to generate training examples for this disease is normally limited to hundreds, even at large cancer centers. To address this relatively sparse data set and evaluate generalizability of the prediction models, we employed cross-validation in algorithm optimization and testing with a new set of samples (collected from different patients and from a later time point) in model validation. In the cross-validation process, the variance of F-score and sensitivity remained small (<0.1), suggesting that the optimized models are generalizable within the sample set. We then collected a separate set of samples, comprising a validation (test) set, which was not used in the training process. With this set, we observed 100% sensitivity, 95% specificity, and an F-score of 0.978 in the optimized SVM model, which values are consistent with the cross-validation results. The results support good generalizability of the optimized prediction models in a larger population. We clarified our work to rule out overfitting in the revised manuscript; the changes are pasted below.

In Page 5,
"Small variances in F-score and sensitivity (<0.1) in cross-validations suggest that the optimized models are generalizable within the sample set.

"To further assess the robustness of the sensor array and algorithm, we synthesized a new batch of OCC-DNAs under the same condition and collected the sensor array response to an independent test set of 54 patient samples (N_{sa} = 54). To evaluate the model performance in various medical conditions, the test set was sampled from different patients, comprised of 7 HGSOC, 5 other gynecologic diseases, 32 non-gynecologic diseases, and 10 heathy patients. With this new sample set, the optimized SVM model resulted in 100% sensitivity at 98% specificity and an F-score of 0.978. These values are consistent with the cross-validation scores and gave a similar receiver operating characteristic (ROC) curve (**Fig 3d**), indicating that the model did not overfit the data."

The work could advance some of the fundamental questions of evaluating current breast cancer biomarkers and possibly even biomarker panels if the authors can elaborate on their model specifically questions 3 and 5. However, the reviewer has concerns as to how this could relate to diagnostic testing. As to the number questions above question 1, 3, 4, and 5 are elements that the reviewer suggests are mandatory to address for further consideration in Nature Biomedical Engineering.

Again, we sincerely appreciate this constructive review. Regarding the reviewer's concern about the relation between the sensor technology and diagnostic testing, we agree that this should be better discussed, and we have modified the manuscript to elaborate on how the sensor technology can benefit the clinical practice and the model performance of our sensor technology compared with current diagnostic testing.

We compared the sensitivity at 98% specificity between our sensor technology and previously reported models with the largest clinical screening trial using the combination of longitudinal CA125 and TVS (Menon *et al Lancet* **2021**). Our optimized model exhibited >87% sensitivity at 98% specificity for HGSOC prediction that is comparable with the largest multi-cohort screening efforts (84% clinical sensitivity).

In terms of potential advantages in clinical applications, there are several unique benefits that we describe here:
1) This method could be rapidly adapted to the detection of many diseases/conditions. The sensor array could be used to train an algorithm to recognize nearly any condition when given enough data from the sensor responses to the appropriate patient serum samples.
2) This method could supplement or replace the use of known biomarkers when there are issues with selectivity. Due to the potential to iteratively modify the sensor array and machine learning algorithms, and to additively augment training set size, the selectivity may be increasingly optimized.
3) This sensor platform can potentially be used in a high-throughput fashion, due to its relatively low intrinsic costs and minimal need for sample processing to facilitate the screening of large patient populations.

4) Because the technology does not rely on antibody-based molecular recognition elements, the sensors could be more robust than existing methods and are not as reliant on cold chain logistics, enabling use in resource-limited settings and in technologies such as point-of-care and wearable/implantable devices.
5) The sensor test will cost much less than the multiplexing of conventional markers judging from a CA125 blood test ($50 or higher per single-analyte test, if the screening accompanied with TVS and/or multiple biomarkers). The low screening cost is also attributed to: 1) the lack of a need for cold chain storage required due to the high physical stability of the sensors, and 2) the small amount of OCC-DNAs needed for screening (<5 nanograms). Currently, the materials needed for the sensor average approximately $5 per sample. The cost of the sensor measurement would also diminish if measured via high-throughput devices, and the potential for the use of very low sample volumes is substantial. To discuss these potential benefits of the technology, we summarized the discussion above in the main text (pasted below).

In Page 8:
"Model performance of the sensor technology exceeds the results of current clinical screening test using longitudinal CA125 and second-line transvaginal ultrasonography (Jacobs *et al Lancet* **2015**) (87% vs. 84% clinical sensitivities at 98% specificity)."

Page 9,
"This sensor technology platform exhibits several unique potential advantages for clinical applications. First, this method could be rapidly adapted to the detection of many diseases/conditions. The array could be used to train an algorithm to recognize nearly any disease when given enough data from the sensor responses to the appropriate patient serum samples. Second, this technology could supplement or replace the use of known biomarkers when there are issues with selectivity in conventional multi-analyte tests. Due to the potential to iteratively modify the sensor array and machine learning algorithms, and to additively augment training set size, the selectivity may be increasingly optimized. Third, this sensor platform can be used in a high-throughput fashion to facilitate the screening of large populations. Fourth, because the technology does not rely on antibody-based molecular recognition elements, the sensors could be more robust than existing methods and are not as reliant on cold chain logistics, enabling use in resource-limited settings and in technologies such as point-of-care and wearable/implantable devices (Lee *et al ACS Sens.* **2019**). Lastly, the sensor technology also has the potential to be developed as an inexpensive and rapid screening tool to result in a single, easy-to-interpret test result in primary care settings. The materials needed for the sensor cost approximately $5 per sample because the lack of need for cold chain storage, the small amount of OCC-DNAs needed for screening (<5 nanograms). The cost of the sensor measurement would also diminish if measured via high-throughput instruments, and the potential for the use of very low sample volumes is substantial."

Reviewer #3 (Report for the authors (Required)):

Brief Summary
This manuscript describes organic color carbon nanotubes, an innovative technology for detecting physicochemical interactions with unknown components of serum from patients. The investigators trained this detection system to differentiate patients with ovarian cancer from patients with other conditions or no condition through analyses of the patients' sera. Six artificial intelligence methods were applied with support vector machines (SVM) having the best performance with over-fitting ameliorated to the extent possible by 10-fold cross validation. The sera are from 269 female patients from six groups: 1. 56 high grade serous ovarian cancer (HGSOC) (most common, most lethal), 2. 71 from patients with normal conditions (although it's unclear why they were patients), 3. 56 with other gynecologic diseases, 4. 61 with non-gynecologic diseases, 5. 7 in HGSOC remission, and 6. 18 in remission in other cancers. The 269 patients are split into a training and validation set (n=215), and a test set (n=54). The SVM approach produced the best result with 95% specificity compared to 81% for CA125, a known serum biomarker for ovarian cancer.

This manuscript describes a significant and original technological advance with carbon nanotubes with potential clinical applications to early detection of ovarian cancer via a blood test. If the test is significantly better than the current standard using CA125, then the impact on early detection of ovarian cancer could be substantial - and there is a clear need for such an improvement in the field. This technology could be applied across the whole field of diagnostic medicine, expanding beyond blood to urine, CSF, and other accessible bodily fluids with the potential of substantial improvements in all diagnostic areas.

This review focusses on the clinical application and does not address the technological advances in carbon nanotubes.

We thank the reviewer for the constructive assessment of the manuscript. We have now thoroughly analyzed the reviewer's comments and address them below.

Major technical criticisms or questions:

1. Criterion for optimization should be appropriate for the clinical application. Early detection of ovarian cancer (and other cancers) requires sensitivity at a high specificity so that benefits outweigh harms. At most 10 surgeries to detect 1 ovarian cancer has been suggested as the limit for ovarian cancer screening. The current approach recently tested in a large screening trial https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00731-5/fulltext uses a rising CA125 with referral to trans-vaginal sonography (TVS). It achieves just under 5 operations to detect 1 ovarian cancer, a 22% positive predictive value (PPV), and 84% of ovarian cancers detected prior to clinical detection (clinical sensitivity). To achieve this overall performance of the screening approach, the blood test has 98% specificity referring 2% of participants to TVS per year, and a sensitivity of 85%. The authors should compare the sensitivity of their blood test at the same specificity (98%) to assess its potential for improving on a rising CA125.

We appreciate the reviewer's comment. As the reviewer suggested, we compared the sensitivity at 98% specificity of our machine learning models with the results of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS) trial (Jacobs *et al Lancet* **2016** and Menon *et al Lancet* **2021**). Our machine learning models can thus far identify the disease with >87% sensitivity at 98% specificity in cross-validation. In the test set, the optimized SVM model resulted in 100% sensitivity at 98% specificity. The model performance exceeds not only the CA125 single threshold rule (58.2% sensitivity; 96.7% specificity – based on the study by Blyuss *et al Clin Cancer Res* **2018**) but also the longitudinal CA125 and second line TVS screening (84% sensitivity at 98% specificity). Additionally, the UKCTOCS algorithm depends on multiple time points while our measurements are based on a single measurement, which is more clinically useful for screening. However, we note that prediction outcomes may differ in clinical screening setting because specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology. Further studies in a high-risk cohort, such as BRCA mutation carriers undergoing risk-reducing surgery, are warranted to evaluate the ability of the model to identify pre-invasive and early-invasive disease. To better evaluate the clinical potential of our sensor technology compared to the current approach, we have reported sensitivity at 98% specificity for our machine learning models in **Table S3** (and pasted at the end of the rebuttal as well) and updated the main text to incorporate these changes.

In Page 5:
"With this new sample set, we observed 100% sensitivity at 98% specificity and an F-score of 0.978 using the SVM model differentiating HGSOC. These values are consistent with the cross-validation scores and gave a similar receiver operating characteristic (ROC) curve (**Fig 3e**), indicating that the model did not overfit the data."

Page 8:
"Model performance of the sensor technology exceeds the results of current best clinical screening test using longitudinal CA125 and second-line transvaginal ultrasonography (Jacobs *et al Lancet* **2015**) (87% vs. 84% clinical sensitivities at 98% specificity). However, considering the fact that

specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, prediction outcomes may differ in clinical screening setting in which specimens are obtained in asymptomatic individuals before clinical diagnosis. Further studies in a high-risk cohort, such as BRCA mutation carriers undergoing risk-reducing surgery, are warranted to evaluate the ability of the model to identify pre-invasive and early-invasive disease."

2. The machine learning literature refers to clinical sensitivity as "recall" = true positives/(true positives + false negatives) = fraction of ovarian cancer cases detected by the test. The other parameter invoked is the precision = true positives/(true positives + false positives) which is the PPV in a clinical context. The authors minimize the F-score = 2/(1/recall + 1/precision) – the harmonic mean of the two parameters, and from a clinical context this criterion is equivalent to 2/(1/sensitivity + 1/PPV) which weights sensitivity and PPV equally. Machine learning has established methods for imbalanced learning so that sensitivity and PPV are not equally weighted. The authors should apply such methods to identify the nanotube test which maximizes sensitivity at 98% specificity = true negative / (true negative + false positive).

We thank the reviewer for this suggestion, which prompted us to conduct additional analysis that is now in the manuscript (**Fig 3d**). To address this point, we performed the Bayesian hyperparameter optimization with the loss function of $(1 - F_\beta)$. The F_β score is the weighted harmonic mean of PPV and sensitivity, $(1+\beta^2)\cdot PPV\cdot sensitivity/(\beta^2\cdot PPV + sensitivity)$ in which β is chosen such that sensitivity is considered β-times as important as PPV. We tested different β values of 0.2, 0.5, 0.8, 1, 1.2, 1.5, 2, and 3 to examine how sensitivity at 98% specificity changes in the prediction models. As β <1 lends more weight to PPV, F_β score weights false positives (related to specificity) more important than true positives (related to sensitivity). We observed that smaller β values systematically increased the sensitivity at 98% specificity, although the improvement was statistically insignificant. The best performing prediction model of the highest clinical sensitivity yielded 87% sensitivity at 98% specificity when PPV and sensitivity were equally weighted (β=1). We have summarized these findings in the figure below for the reviewer, and the plot and optimization results are included as panel **d** in **Fig 3** (and pasted below) and **Table S3**. A new discussion of this issue is now provided in the text and pasted below.

Page 5
"Lastly, we examined if tuning the hyperparameters to maximize the F_β score can improve sensitivity at a high specificity (**Fig 3d**). The F_β score is the weighted harmonic mean of PPV and sensitivity and β is chosen such that sensitivity is considered β-times as important as PPV. At decreasing β from 3 to 0.2, sensitivity at 98% specificity systematically increased, although the improvement was statistically insignificant. The best performing prediction model was the sensor array combination of 4-$N(C_2H_5)_2$*$CT_2C_3T_2C$, 4-$N(C_2H_5)_2$*$(TAT)_4$, 3,4,5-F_3*$(TAT)_4$, 3,4,5-F_3*$(AT)_{15}$, and 3,4,5-F_3*$(GT)_{15}$, and yielded 87% sensitivity at 98% specificity when PPV and sensitivity were equally weighted (β =1)."

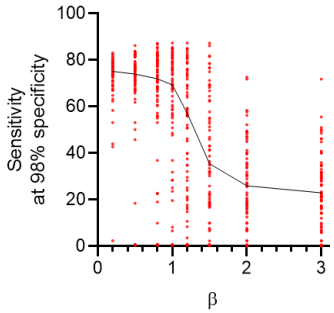

Fig. 3 | Optimization of machine learning algorithms for HGSOC classification. … **d,** Sensitivity at 98% specificity obtained with varying βin F_β scoring via SVM. The line connects the median of sensitivities for the optimized nanosensor arrays. …

3. The abstract states that ovarian cancer is challenging to detect because biomarkers are elevated in

other conditions, and that most biomarkers are not specific enough for screening purposes. This is not true. A rising CA125 followed by TVS for a positive blood test has a specificity of 99.8% and a PPV exceeding 20%, twice the minimum suggested in the literature. Instead, the challenge is to maximize sensitivity for early-stage disease while maintaining the specificity at an achievable 98%.

We thank the reviewer for correcting our statement. We revised the text associated with challenges in ovarian cancer screening. We are aware that the challenge in early-stage detection is to maximize sensitivity at high specificity. In the abstract, we wanted to note that biomarker-based screening (e.g. CA125 single threshold) is not specific enough for representing the whole disease state. We modified Abstract to avoid confusion and clarified what the main challenge in ovarian cancer detection is in Introduction.

Abstract:
"Serum biomarkers are widely used as diagnostic indicators, but many are not specific and/or sensitive enough for screening purposes. Ovarian cancer, for example, is challenging to diagnose, in part because the biomarker levels are elevated in other conditions."

In Page 1,
"Conventionally, serum biomarker measurements, such as cancer antigen 125 (CA125) combined with transvaginal ultrasonography, have been suggested for use as a screening tool for the detection of ovarian cancer (Menon *et al J Clin Oncol* **2015** and Blyuss *et al Clin Cancer Res* **2018**). Recent reports have found that these methods do not result in early-stage detection and confer little survival benefit (Jacobs *et al Lancet* **2016**, Menon *et al Lancet* **2021**) in part due to the challenge of improving sensitivity while maintaining high specificity. Other complementary serum biomarkers such as human epidermis protein 4 (HE4), chitinase-3-like protein 1 (YKL40), and mesothelin, or panels of biomarkers have been reported to result in higher sensitivity over CA125-based screening. However, the improvement in discriminatory power for ovarian cancer diagnosis is still under debate. Currently, no screening strategy has been shown to reduce mortality, and screening strategies are associated with a high rate of false-positive results and a risk of harm from invasive testing."

4. The abstract states that their strategy yields a specificity of 95% while CA125 has a specificity of 81%. These specificity values depend on which cutoff is chosen. To compare to the best method applied in clinical screening trials, a rising CA125 followed by TVS, the authors' methods need to be compared to this first line blood test which has a specificity of 98% and a sensitivity of 84%.

We appreciate the reviewer's comment. We now compare the sensitivity at 98% specificity of the nanosensor array with the CA125-triggered TVS test results (Menon *et al Lancet* **2021**). The text in the abstract now reads:

Abstract:
"This strategy yielded 87% sensitivity at 98% specificity, versus 84% via the multimodal test using the biomarker, cancer antigen 125 and transvaginal ultrasonography."

5. Please explain the choice of other conditions since such conditions are rare in a general population screening context. If the aim is to develop a blood test for early detection of ovarian cancer, should not the "healthy controls" form the majority of the control group?

We agree with the reviewer that 'healthy controls' should be the majority of the control group in a general population screening context. We developed the prediction model to differentiate HGSOC not only from healthy controls but also from patients with other diseases. We reasoned that the sensor needs to be effective in both environments. Because we were not logistically able to expand the number of patient samples by a large amount in the initial study, we chose to ask whether the nanosensor/algorithm response would be likely to differentiate the response from other conditions where known biomarkers are likely to produce false-positive responses.

Minor technical questions:
1. Presumably cases were HGSOC while all other conditions were classified as “controls”. Please clarify.

We thank the reviewer for correcting the error. We clarified all other conditions, including non-gynecologic disease and healthy patients as controls. We made the correction in the manuscript.

2. The text states that “a single subsample was used to test the trained model”. Please clarify what testing in this context means, and why n=54 was chosen for the size of the test set.

In Methods, we have clarified how we performed the cross-validation. The text describing model validation now reads:

Model training and performance assessment in Methods (Page 11):
“In the cross-validation process, stratified shuffle split validation was used to randomly partition the data set into ten subsamples. Each partition, nine of the ten subsamples were used to train the model while a single subsample was used to calculate the F-score of the trained model. The average F-score of the ten-fold cross-validation was used to assess model performance.”

Regarding the size of test set (n=54), we included 7 HGSOC, 5 other gynecologic diseases, 32 non-gynecologic diseases, and 10 heathy patients to evaluate the model performance in various medical conditions. The test set was sampled from different patients who were not included in the training and cross-validation sets. We clarified how we sampled and organized the test set in the manuscript.

In Page 5:
“To further assess the robustness of the sensor array and algorithm, we synthesized a new batch of OCC-DNAs under the same condition and collected the sensor array response to an independent test set of 54 patient samples (N_{sa} = 54). To evaluate the model performance in various medical conditions, the test set was sampled from different patients, comprised of 7 HGSOC, 5 other gynecologic diseases, 32 non-gynecologic diseases, and 10 heathy patients.”

Model training and performance assessment in Methods (Page 11):
“The trained models were then tested with an independent set of patient sera (N=54), sampled from different patients (test set), as external validation.”

3. The authors should define “recall” and “precision” and F-score for clinical readers in terms of true positives, false positives, true negatives, and false negatives, and relate these terms to “clinical sensitivity” and “positive predictive value”.

We thank the reviewer’s suggestion. Now we replaced recall as sensitivity and precision as positive predictive value (PPV) throughout the paper. F-score is defined as the harmonic means of clinical sensitivity and PPV.

4. Missing citations to relevant literature
The authors should refer to the largest ovarian cancer screening trial:
1. https://pubmed.ncbi.nlm.nih.gov/26707054/, or
2. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)00731-5/fulltext

We are grateful for the reviewer’s insight into the field and have added these relevant references in the text as reference 5 (Jacobs *et al Lancet* **2016**) and 6 (Menon *et al Lancet* **2021**). The specific lines that refer to these references now read:

Page 1:

“Conventionally, serum biomarker measurements, such as cancer antigen 125 (CA125) combined with transvaginal ultrasonography, have been suggested for use as a screening tool for the detection of ovarian cancer (Menon *et al J Clin Oncol* **2015**, Blyuss *et al Clin Cancer Res* **2018**). Recent reports have found that these methods do not result in early-stage detection and confer little survival benefit (Jacobs *et al Lancet* **2016**, Menon *et al Lancet* **2021**) in part due to the challenge of improving sensitivity while maintaining high specificity.”

Page 8:
“Model performance of the sensor technology exceeds the results of the current clinical test using longitudinal CA125 and second-line transvaginal ultrasonography (Jacobs *et al Lancet* **2015**) (87% vs. 84% clinical sensitivities at 98% specificity), and the technology may enable reporting faster and more accurate test results to clinicians.”

Appendix 1

PROBAST (**P**rediction model study **R**isk **O**f **B**ias **AS**sessment **T**ool)
Version of 15/05/2019 – Page 24
For more information, please see www.probast.org

Step 1: Specify your systematic review question

Criteria	**Specify your systematic review question**
Intended use of model:	Diagnosis of high-grade serous ovarian cancer using a machine perception nanosensor array
***Participants** including selection criteria and setting:*	Female adults patients diagnosed with ovarian cancer or other diseases, and healthy female blood donors
***Predictors**:*	Near-infrared fluorescence response of carbon nanotube based sensors
Outcome to be predicted:	High-grade serous ovarian cancer

Step 2: Classify the type of prediction model evaluation

Classify the evaluation based on its aim			
Type of prediction study	**PROBAST boxes to complete**	**Tick as appropriate**	**Definition for type of prediction model study**
Development only	Development	X	Prediction model development without external validation. These studies may include internal validation methods, such as bootstrapping and cross-validation techniques.
Development and validation	Development and validation	✓	Prediction model development combined with external validation in other participants in the same article.
Validation only	Validation	X	External validation of existing (previously developed) model in other participants.

Step 3: Assess risk of bias and applicability

DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection: Fluorescence responses of nanosensors in patient serum were collected using a custom-built near-infrared plate reader. Patient demographics (age and sex), medication, hysterectomy, diagnosis, and biomarker levels were acquired by chart review.			
		Dev	Val
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?		Y	Y
1.2 Were all inclusions and exclusions of participants appropriate?		PN	PN
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*	**high**	**high**
Rationale of bias rating: Because the medical record of healthy donors was not available, it is possible that selection bias was introduced to the healthy control.			

A large fraction of breast cancers in the non-HGSOC group of the test set may introduce systematic bias.
Specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of our technology.

B. Applicability

Describe included participants, setting and dates:
Serum samples were collected from 269 patients between August 2018 and March 2021 in the Memorial Sloan Kettering Cancer Center, USA.

Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**low**	**low**

Rationale of applicability rating:
Included patients appear representative of the population specified in the review question

DOMAIN 2: Predictors

A. Risk of Bias

List and describe predictors included in the final model, e.g. definition and timing of assessment:

Predictors were defined as a difference in sensor response acquired from patient serum and phosphate buffered silane (PBS). Specifically, the E_{11} peak position feature, *dwl*, was defined as the wavelength difference between the E_{11} peak in the patient sample, wl, and PBS, wl_0, *dwl* = wl – wl_0. The E_{11} peak intensity feature, *dint*, was normalized as *dint* = int/int_0, where int and int_0 are the E_{11} peak intensity in serum and PBS, respectively. Similarly, we defined E_{11}^- peak related features, *dwl** and *dint**, indicating the relative E_{11}^- peak position and intensity.
The nanosensors were incubated in 20% serum at room temperature for 2 hours and in a cold room (4 ^oC) after the spectral acquisition at 2-hour time point. Data were taken at three time points during incubation: 2 hours, 24 hours, and 72 hours.

		Dev	Val
2.1 Were predictors defined and assessed in a similar way for all participants?		Y	Y
2.2 Were predictor assessments made without knowledge of outcome data?		Y	Y
2.3 Are all predictors available at the time the model is intended to be used?		Y	Y
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*	**low**	**low**

Rationale of bias rating:
They appear to match the review question.

B. Applicability

Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**low**	**low**

Rationale of applicability rating:
They appear to match the review question.

DOMAIN 3: Outcome

A. Risk of Bias

Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:

Diagnoses were identified from a chart review of each patient; all diagnoses included histology and were confirmed by a gynecologic oncology attending physician.
There is no time interval between predictor assessment (spectral acquisition at 2, 24, and 72 hours of incubation) and outcome determination.

		Dev	Val
3.1 Was the outcome determined appropriately?		Y	Y
3.2 Was a pre-specified or standard outcome definition used?		Y	Y
3.3 Were predictors excluded from the outcome definition?		Y	Y
3.4 Was the outcome defined and determined in a similar way for all participants?		Y	Y
3.5 Was the outcome determined without knowledge of predictor information?		Y	Y
3.6 Was the time interval between predictor assessment and outcome determination appropriate?		Y	Y
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*	**low**	**low**
Rationale of bias rating: They appear to match the review question.			
B. Applicability			
At what time point was the outcome determined: Outcome was determined at the same moment as the measurement of the predictors which is in line with the diagnostic nature of the study *If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:* N/A			
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*	**low**	**low**
Rationale of applicability rating: The outcome is the same as that specified in the review question.			

DOMAIN 4: Analysis
Risk of Bias
Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor: Participants: 269 waste samples were collected from female healthy blood donors and patients diagnosed with ovarian and other cancers under a Memorial Sloan Kettering Cancer Center Institutional Review Board approved protocol. From this sample set, 56 specimens were collected from patients diagnosed with high grade serous ovarian cancer, 71 specimens from healthy donors, 56 with other gynecologic diseases, 61 with non-gynecologic diseases, and 25 in remission. Candidate predictors: 24 = 6 nanosensor types * 4 different spectral parameters Outcome events: 2, high-grade serous ovarian cancer vs. others Number of events per candidate predictor: unclear
Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition): We created models based on Decision Tree, Logistic Regression, Artificial Neural Networks, Random Forest, and Support Vector Machine for binary classification. Hyperparameters for each model were optimized using Bayesian Optimization, implemented in the HyperOpt library. The loss function to minimize in the hyperparameter optimization was set to (1 – F-score). F-score is a measure of accuracy in binary classification and calculated from the harmonic mean of PPV and sensitivity. To identify the optimal predictor combinations, we optimized the machine learning algorithms at varying sensor combinations, durations of incubation, and spectral variables within the sensor array.
Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants):

Model performance was evaluated using ten-fold cross-validation. In the cross-validation process, stratified shuffle split validation was used to randomly partition the data set into ten subsamples. Each partition, nine of the ten subsamples were used to train the model while a single subsample was used to test the trained model. The average F-score of the ten-fold cross-validation was used to assess model performance. The trained models were then tested with an independent set of patient sera, sampled from different patients and from a later time point (test set) as an external validation.

Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism:
The averaged F-score and clinical sensitivity (sensitivity at 98% specificity) were used to determine the model performance.

Describe any participants who were excluded from the analysis:
N/A

Describe missing data on predictors and outcomes as well as methods used for missing data:
N/A

		Dev	Val
4.1 Were there a reasonable number of participants with the outcome?		PY	PY
4.2 Were continuous and categorical predictors handled appropriately?		NI	NI
4.3 Were all enrolled participants included in the analysis?		Y	Y
4.4 Were participants with missing data handled appropriately?		NI	NI
4.5 Was selection of predictors based on univariable analysis avoided?		Y	
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?		PY	Y
4.7 Were relevant model performance measures evaluated appropriately?		Y	Y
4.8 Were model overfitting and optimism in model performance accounted for?		Y	
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?		NI	
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*	**unclear**	**low**

Rationale of bias rating:
Background chronic diseases could affect the blood constituents, introducing interferences in sensor response. Chemical and biological interferents could interact with the nanosensors, contributing to the fluorescence response. From a patient chart review, we identified chronic diseases and most-common medications administered to the patients. We found that 75% of HGSOC and 68% of other disease patients suffered from at least one chronic condition, and the relative abundance between these disease groups was similar. Regarding the medications, we statistically assessed the contribution of each interferent to the sensor results using a multivariate regression model. The regression model determined a linear correlation between the sensor response and medication, using estimated parameters and errors. The adjusted R^2 of the regression model ranged from -0.045 to 0.233, indicating weak linear correlations of each sensor response to medications. We confirmed that the sensor array platform accurately classified the disease status regardless of medication and chronic conditions, evidenced by high F-scores of the HGSOC prediction models. The analysis suggested no indications that such interferences reduced specificity in HGSOC detection by the sensor platform.

Step 4: Overall assessment

Overall judgement about risk of bias and applicability of the prediction model evaluation		
Overall judgement of risk of bias	**RISK:** *(low/ high/ unclear)*	**high**
Summary of sources of potential bias: In terms of participants, the medical record of healthy donors was not available, potentially introducing selection bias to the non-ovarian cancer group. A large fraction of breast cancers in the non-HGSOC group of the test set may introduce systematic bias in participant selection and the development of machine learning models. Specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of our technology. Comorbidities such as background chronic diseases could affect the blood constituents. Chemical and biological interferents could interact with the nanosensors, introducing interferences in sensor response.		
Overall judgement of applicability	**CONCERN:** *(low/ high/ unclear)*	**low**
Summary of applicability concerns: No concerns		

Table S1 | Model performance of all machine learning algorithms and OCC-DNA sensor array combinations tested for binary HGSOC classification using the training and validation set of samples (N_{sa}=215). OCC-DNA sensor responses of 2-hour incubation were used if not specified. The F1-score, positive predictive value (PPV), sensitivity, and sensitivity at 98% specificity were the averaged values in 10-fold cross-validation. The spectral variables used to train each algorithm are specified below. Abbreviations: Logi = Logistic regression, DT = Decision tree, ANN = Artificial neural network, RF = Random forest, SVM = Support vector machine.

(1) Two spectra variables, *Δwl* + *Δint* (N_f = 2):

OCC-DNA in the sensor array (2 variables)	F1-score					PPV					Sensitivity					Sensitivity at 98% specificity				
	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM
4-N$(C_2H_5)_2$*$CT_2C_3T_2C$	0.667	0.658	0.695	0.746	0.745	0.594	0.600	0.555	0.697	0.666	0.765	0.741	0.940	0.807	0.851	0.096	0.033	0.055	0.243	0.142
4-N$(C_2H_5)_2$*$(TAT)_4$	0.714	0.774	0.705	0.754	0.761	0.732	0.717	0.711	0.716	0.707	0.705	0.848	0.704	0.800	0.830	0.170	0.177	0.175	0.168	0.102
3,4,5-F_3*$CT_2C_3T_2C$	0.667	0.693	0.676	0.707	0.744	0.500	0.532	0.511	0.651	0.732	1.000	0.994	1.000	0.782	0.765	0.006	0.023	0.060	0.240	0.113
3,4,5-F_3*$(TAT)_4$	0.721	0.728	0.729	0.775	0.768	0.738	0.707	0.760	0.759	0.761	0.711	0.760	0.705	0.796	0.790	0.245	0.240	0.264	0.390	0.127
3,4,5-F_3*$(AT)_{15}$	0.631	0.737	0.691	0.778	0.779	0.655	0.675	0.535	0.731	0.707	0.620	0.819	0.976	0.836	0.874	0.122	0.119	0.000	0.207	0.084
3,4,5-F_3*$(GT)_{15}$	0.739	0.747	0.807	0.831	0.827	0.694	0.735	0.728	0.803	0.764	0.796	0.760	0.910	0.867	0.903	0.169	0.052	0.188	0.214	0.187
4-N$(C_2H_5)_2$*$CT_2C_3T_2C$ + 4-N$(C_2H_5)_2$*$(TAT)_4$	0.686	0.776	0.753	0.850	0.858	0.624	0.725	0.698	0.784	0.815	0.772	0.843	0.818	0.933	0.909	0.037	0.232	0.180	0.253	0.096
4-N$(C_2H_5)_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.675	0.733	0.692	0.791	0.816	0.598	0.640	0.635	0.755	0.834	0.777	0.878	0.771	0.836	0.807	0.095	0.041	0.089	0.445	0.469
4-N$(C_2H_5)_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.716	0.777	0.778	0.863	0.879	0.695	0.763	0.718	0.830	0.827	0.746	0.801	0.856	0.903	0.940	0.125	0.238	0.310	0.618	0.167
4-N$(C_2H_5)_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.682	0.787	0.698	0.821	0.823	0.644	0.749	0.655	0.777	0.806	0.735	0.836	0.753	0.873	0.843	0.062	0.166	0.103	0.343	0.372
4-N$(C_2H_5)_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.743	0.824	0.743	0.882	0.885	0.700	0.787	0.701	0.832	0.839	0.796	0.867	0.796	0.940	0.940	0.188	0.209	0.182	0.474	0.307
4-N$(C_2H_5)_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.703	0.748	0.717	0.835	0.887	0.722	0.742	0.573	0.809	0.844	0.692	0.759	0.964	0.871	0.940	0.168	0.165	0.179	0.424	0.372
4-N$(C_2H_5)_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$	0.719	0.750	0.754	0.790	0.840	0.724	0.672	0.715	0.759	0.804	0.722	0.855	0.806	0.830	0.885	0.193	0.088	0.283	0.403	0.111
4-N$(C_2H_5)_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.719	0.766	0.758	0.804	0.860	0.721	0.690	0.716	0.764	0.870	0.722	0.866	0.812	0.854	0.860	0.157	0.075	0.204	0.336	0.620
4-N$(C_2H_5)_2$*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.788	0.823	0.794	0.841	0.863	0.741	0.773	0.749	0.804	0.803	0.844	0.886	0.850	0.890	0.939	0.252	0.257	0.258	0.522	0.085
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.723	0.788	0.764	0.842	0.860	0.698	0.745	0.769	0.824	0.833	0.758	0.843	0.765	0.867	0.898	0.146	0.295	0.302	0.490	0.264
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.663	0.761	0.700	0.824	0.837	0.650	0.728	0.557	0.780	0.775	0.688	0.806	0.947	0.879	0.916	0.054	0.114	0.000	0.259	0.185
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.742	0.821	0.744	0.857	0.870	0.697	0.754	0.701	0.802	0.813	0.796	0.904	0.796	0.921	0.940	0.182	0.137	0.126	0.418	0.154
3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.723	0.714	0.758	0.819	0.846	0.712	0.711	0.750	0.784	0.784	0.740	0.723	0.777	0.861	0.922	0.132	0.154	0.282	0.464	0.187
3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.766	0.802	0.811	0.858	0.864	0.735	0.767	0.737	0.832	0.819	0.801	0.844	0.905	0.892	0.916	0.296	0.371	0.233	0.573	0.322
3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.740	0.795	0.746	0.830	0.865	0.686	0.748	0.708	0.798	0.843	0.808	0.853	0.796	0.872	0.896	0.145	0.133	0.126	0.433	0.346
4-N$(C_2H_5)_2$*$CT_2C_3T_2C$ + 4-N$(C_2H_5)_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.763	0.804	0.795	0.842	0.890	0.737	0.748	0.761	0.798	0.850	0.796	0.874	0.838	0.898	0.940	0.269	0.240	0.282	0.614	0.222

OCC-DNA in the sensor array (2 variables)	F1-score					PPV					Sensitivity					Sensitivity at 98% specificity				
	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$	0.743	0.797	0.787	0.862	0.868	0.701	0.701	0.731	0.812	0.834	0.796	0.928	0.861	0.921	0.910	0.188	0.193	0.146	0.518	0.458
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.767	0.843	0.838	0.858	0.893	0.733	0.806	0.774	0.819	0.879	0.808	0.891	0.917	0.903	0.909	0.313	0.267	0.320	0.605	0.490
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.732	0.738	0.777	0.854	0.861	0.738	0.725	0.734	0.824	0.836	0.734	0.765	0.832	0.891	0.892	0.181	0.185	0.257	0.591	0.483
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.780	0.807	0.807	0.846	0.899	0.736	0.746	0.771	0.789	0.881	0.832	0.886	0.849	0.915	0.921	0.228	0.226	0.193	0.390	0.455
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.801	0.811	0.823	0.854	0.900	0.769	0.740	0.757	0.820	0.883	0.838	0.904	0.904	0.897	0.922	0.281	0.190	0.251	0.490	0.563
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.725	0.801	0.806	0.820	0.896	0.712	0.735	0.746	0.787	0.884	0.746	0.885	0.886	0.866	0.914	0.180	0.161	0.215	0.509	0.701
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.780	0.818	0.792	0.872	0.895	0.732	0.759	0.745	0.817	0.925	0.838	0.892	0.850	0.940	0.872	0.251	0.158	0.271	0.566	0.788
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.724	0.788	0.788	0.866	0.915	0.720	0.743	0.723	0.807	0.918	0.733	0.854	0.874	0.939	0.916	0.163	0.126	0.133	0.463	0.796
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.736	0.764	0.779	0.838	0.908	0.726	0.691	0.723	0.788	0.887	0.753	0.861	0.849	0.903	0.933	0.169	0.111	0.336	0.500	0.649
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.761	0.835	0.818	0.877	0.913	0.749	0.799	0.762	0.818	0.921	0.777	0.879	0.886	0.945	0.908	0.174	0.268	0.160	0.532	0.636
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.795	0.821	0.831	0.876	0.906	0.761	0.752	0.762	0.838	0.882	0.839	0.910	0.916	0.921	0.933	0.295	0.245	0.270	0.670	0.616
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.729	0.755	0.780	0.851	0.869	0.712	0.752	0.736	0.813	0.820	0.752	0.778	0.838	0.897	0.928	0.185	0.225	0.390	0.566	0.220
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.745	0.826	0.750	0.875	0.886	0.704	0.758	0.688	0.816	0.837	0.796	0.910	0.831	0.946	0.946	0.200	0.101	0.274	0.500	0.520
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.683	0.750	0.702	0.832	0.842	0.595	0.650	0.688	0.792	0.788	0.808	0.899	0.723	0.879	0.909	0.043	0.038	0.092	0.344	0.049
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.710	0.770	0.758	0.849	0.887	0.700	0.722	0.774	0.802	0.840	0.729	0.831	0.748	0.909	0.945	0.277	0.148	0.345	0.536	0.206
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.766	0.804	0.789	0.873	0.918	0.737	0.766	0.741	0.827	0.909	0.801	0.849	0.850	0.927	0.933	0.234	0.142	0.228	0.581	0.728
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.732	0.802	0.767	0.857	0.892	0.730	0.753	0.724	0.803	0.887	0.740	0.866	0.820	0.927	0.903	0.217	0.146	0.132	0.301	0.694
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.745	0.782	0.795	0.827	0.883	0.734	0.737	0.733	0.757	0.862	0.765	0.842	0.873	0.914	0.910	0.188	0.063	0.234	0.546	0.586
3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.725	0.781	0.723	0.830	0.876	0.722	0.729	0.727	0.782	0.827	0.734	0.855	0.728	0.891	0.934	0.186	0.209	0.204	0.540	0.372
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$	0.757	0.796	0.794	0.853	0.892	0.726	0.756	0.753	0.823	0.907	0.796	0.850	0.844	0.892	0.885	0.251	0.169	0.282	0.601	0.585

OCC-DNA in the sensor array (2 variables)	F1-score					PPV					Sensitivity					Sensitivity at 98% specificity				
	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.785	0.812	0.821	0.853	0.912	0.768	0.760	0.754	0.811	0.891	0.807	0.879	0.904	0.908	0.939	0.282	0.148	0.084	0.546	0.575
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.767	0.805	0.813	0.874	0.906	0.724	0.772	0.757	0.809	0.900	0.820	0.849	0.879	0.952	0.915	0.228	0.286	0.206	0.472	0.750
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.807	0.820	0.819	0.871	0.914	0.777	0.775	0.764	0.827	0.892	0.844	0.879	0.886	0.928	0.939	0.306	0.177	0.263	0.603	0.760
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.732	0.803	0.801	0.858	0.912	0.720	0.748	0.743	0.810	0.913	0.752	0.872	0.874	0.920	0.915	0.174	0.195	0.227	0.495	0.843
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.756	0.812	0.813	0.866	0.907	0.728	0.733	0.766	0.823	0.857	0.789	0.916	0.873	0.916	0.969	0.253	0.136	0.290	0.682	0.272
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$	0.754	0.837	0.806	0.880	0.889	0.741	0.788	0.738	0.823	0.880	0.772	0.897	0.891	0.946	0.903	0.169	0.331	0.159	0.518	0.764
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.756	0.803	0.795	0.859	0.908	0.742	0.773	0.752	0.810	0.899	0.777	0.843	0.850	0.915	0.921	0.320	0.282	0.338	0.636	0.759
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.732	0.751	0.775	0.869	0.902	0.726	0.717	0.746	0.832	0.856	0.746	0.794	0.819	0.916	0.958	0.186	0.101	0.313	0.607	0.421
No 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.771	0.823	0.830	0.871	0.929	0.750	0.767	0.758	0.815	0.911	0.795	0.890	0.922	0.939	0.951	0.217	0.124	0.240	0.564	0.687
No 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.802	0.838	0.817	0.868	0.922	0.766	0.805	0.768	0.812	0.892	0.844	0.880	0.874	0.939	0.957	0.264	0.297	0.287	0.656	0.564
No 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.726	0.808	0.815	0.861	0.910	0.721	0.763	0.784	0.827	0.906	0.740	0.867	0.856	0.903	0.921	0.174	0.124	0.270	0.607	0.744
No 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.768	0.833	0.816	0.861	0.913	0.736	0.775	0.760	0.808	0.897	0.807	0.911	0.885	0.927	0.933	0.247	0.170	0.247	0.480	0.770
No 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.734	0.780	0.777	0.860	0.920	0.725	0.721	0.734	0.834	0.909	0.753	0.861	0.831	0.896	0.934	0.217	0.192	0.157	0.449	0.837
No 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.737	0.793	0.772	0.853	0.893	0.732	0.752	0.735	0.785	0.892	0.753	0.849	0.819	0.939	0.897	0.170	0.200	0.245	0.522	0.714
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$	0.785	0.803	0.811	0.884	0.914	0.762	0.762	0.783	0.822	0.883	0.814	0.856	0.844	0.958	0.951	0.282	0.108	0.258	0.559	0.696
No 4-N(C_2H_5)$_2$*$(TAT)_4$	0.762	0.803	0.823	0.880	0.902	0.734	0.748	0.757	0.838	0.846	0.795	0.873	0.904	0.928	0.969	0.258	0.250	0.283	0.683	0.499
No 3,4,5-F_3*$CT_2C_3T_2C$	0.785	0.812	0.802	0.866	0.925	0.752	0.770	0.755	0.819	0.908	0.826	0.867	0.856	0.922	0.945	0.252	0.220	0.275	0.656	0.749
No 3,4,5-F_3*$(TAT)_4$	0.771	0.827	0.829	0.898	0.928	0.734	0.776	0.760	0.842	0.916	0.813	0.892	0.916	0.964	0.946	0.187	0.175	0.200	0.580	0.866
No 3,4,5-F_3*$(AT)_{15}$	0.806	0.828	0.821	0.879	0.930	0.772	0.760	0.768	0.831	0.905	0.850	0.915	0.885	0.939	0.957	0.324	0.154	0.299	0.639	0.760
No 3,4,5-F_3*$(GT)_{15}$	0.728	0.787	0.794	0.871	0.915	0.726	0.738	0.742	0.836	0.919	0.740	0.856	0.861	0.915	0.915	0.185	0.072	0.225	0.583	0.867
All six OCC-DNA sensors	0.791	0.804	0.820	0.888	0.916	0.758	0.781	0.761	0.829	0.902	0.832	0.842	0.892	0.958	0.933	0.258	0.194	0.252	0.628	0.847
All six OCC-DNA sensors with 24h incubation	0.744	0.784	0.820	0.841	0.882	0.707	0.711	0.758	0.781	0.861	0.790	0.879	0.898	0.915	0.909	0.189	0.058	0.197	0.560	0.464
All six OCC-DNA sensors with 72h incubation	0.736	0.809	0.851	0.894	0.907	0.665	0.725	0.788	0.846	0.903	0.831	0.922	0.934	0.952	0.916	0.122	0.080	0.289	0.559	0.723
All six OCC-DNA sensors with 2 h and 24 h incubation	0.801	0.833	0.862	0.884	0.932	0.759	0.780	0.803	0.838	0.960	0.850	0.898	0.933	0.940	0.909	0.235	0.183	0.362	0.561	0.861
All six OCC-DNA sensors with 2 h, 24 h, and 72 h incubation	0.817	0.819	0.913	0.895	0.931	0.804	0.811	0.870	0.847	0.938	0.837	0.837	0.964	0.951	0.928	0.289	0.189	0.444	0.597	0.886

(2) Four spectra variables, *dwl* + *dwl** + *dint* + *dint** (N_f = 4):

OCC-DNA in the sensor array (4 variables)	F1-score					PPV					Sensitivity					Sensitivity at 98% specificity				
	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$	0.728	0.822	0.729	0.879	0.929	0.691	0.789	0.668	0.838	0.917	0.778	0.867	0.808	0.927	0.945	0.072	0.199	0.048	0.449	0.768
4-N(C_2H_5)$_2$*$(TAT)_4$	0.727	0.812	0.794	0.848	0.874	0.665	0.724	0.749	0.799	0.841	0.808	0.934	0.849	0.909	0.915	0.120	0.160	0.167	0.398	0.237
3,4,5-F_3*$CT_2C_3T_2C$	0.728	0.811	0.758	0.822	0.880	0.671	0.739	0.699	0.770	0.840	0.802	0.904	0.832	0.890	0.927	0.102	0.102	0.079	0.357	0.487
3,4,5-F_3*$(TAT)_4$	0.773	0.850	0.804	0.889	0.913	0.731	0.783	0.729	0.848	0.915	0.826	0.933	0.904	0.940	0.915	0.217	0.275	0.105	0.593	0.785
3,4,5-F_3*$(AT)_{15}$	0.741	0.820	0.808	0.859	0.909	0.689	0.771	0.730	0.792	0.905	0.808	0.879	0.910	0.939	0.914	0.065	0.166	0.156	0.479	0.748
3,4,5-F_3*$(GT)_{15}$	0.733	0.835	0.731	0.874	0.904	0.675	0.790	0.671	0.835	0.903	0.808	0.891	0.808	0.921	0.909	0.126	0.099	0.035	0.496	0.587
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$	0.749	0.806	0.833	0.861	0.912	0.706	0.721	0.782	0.817	0.872	0.802	0.915	0.898	0.915	0.958	0.210	0.092	0.303	0.512	0.252
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.731	0.836	0.805	0.880	0.911	0.671	0.773	0.742	0.841	0.912	0.808	0.917	0.886	0.928	0.915	0.150	0.159	0.250	0.614	0.681
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.732	0.783	0.758	0.844	0.890	0.649	0.706	0.684	0.784	0.892	0.844	0.885	0.856	0.920	0.896	0.078	0.093	0.090	0.451	0.694
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.766	0.816	0.817	0.874	0.930	0.728	0.753	0.757	0.825	0.935	0.815	0.897	0.893	0.933	0.927	0.196	0.118	0.246	0.572	0.781
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.757	0.757	0.760	0.803	0.834	0.637	0.683	0.644	0.751	0.793	0.940	0.855	0.934	0.872	0.885	0.072	0.043	0.097	0.365	0.067
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.760	0.829	0.823	0.870	0.897	0.701	0.762	0.772	0.817	0.856	0.838	0.917	0.892	0.933	0.945	0.211	0.093	0.211	0.503	0.509
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$	0.770	0.809	0.844	0.862	0.906	0.698	0.735	0.777	0.810	0.938	0.862	0.904	0.927	0.927	0.879	0.223	0.141	0.276	0.562	0.695
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.727	0.787	0.769	0.836	0.877	0.665	0.710	0.671	0.806	0.909	0.808	0.885	0.909	0.878	0.853	0.053	0.102	0.108	0.456	0.646
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.729	0.808	0.776	0.863	0.889	0.669	0.759	0.749	0.837	0.841	0.808	0.874	0.807	0.897	0.946	0.060	0.200	0.250	0.558	0.197
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.727	0.830	0.759	0.853	0.895	0.665	0.779	0.701	0.787	0.894	0.808	0.892	0.832	0.939	0.897	0.054	0.197	0.204	0.337	0.677
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.789	0.811	0.817	0.855	0.918	0.717	0.724	0.728	0.794	0.916	0.880	0.928	0.940	0.927	0.921	0.205	0.081	0.189	0.474	0.759
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.773	0.840	0.808	0.869	0.901	0.740	0.783	0.723	0.815	0.865	0.814	0.910	0.922	0.934	0.946	0.190	0.233	0.211	0.474	0.624
3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.735	0.832	0.776	0.859	0.873	0.678	0.770	0.703	0.806	0.823	0.808	0.909	0.874	0.922	0.934	0.103	0.151	0.186	0.524	0.285
3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.726	0.810	0.727	0.868	0.903	0.677	0.774	0.664	0.842	0.865	0.790	0.856	0.808	0.898	0.946	0.060	0.125	0.108	0.574	0.478
3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.727	0.800	0.756	0.832	0.865	0.665	0.693	0.672	0.781	0.846	0.808	0.952	0.868	0.898	0.892	0.047	0.169	0.137	0.225	0.599
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.799	0.813	0.844	0.868	0.913	0.756	0.707	0.815	0.814	0.918	0.850	0.964	0.879	0.934	0.910	0.127	0.141	0.307	0.511	0.796
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$	0.787	0.848	0.833	0.885	0.914	0.731	0.793	0.780	0.845	0.945	0.856	0.922	0.904	0.933	0.890	0.199	0.196	0.274	0.504	0.806
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.767	0.816	0.859	0.866	0.919	0.746	0.747	0.823	0.811	0.949	0.794	0.909	0.910	0.933	0.892	0.144	0.106	0.412	0.529	0.825
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.757	0.818	0.802	0.852	0.895	0.692	0.712	0.760	0.793	0.919	0.838	0.964	0.856	0.927	0.878	0.096	0.069	0.232	0.571	0.763

OCC-DNA in the sensor array (4 variables)	F1-score					PPV					Sensitivity					Sensitivity at 98% specificity				
	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.734	0.818	0.796	0.870	0.900	0.648	0.749	0.735	0.820	0.912	0.850	0.921	0.873	0.933	0.890	0.199	0.227	0.266	0.564	0.793
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.807	0.859	0.817	0.882	0.924	0.742	0.807	0.768	0.829	0.940	0.886	0.921	0.874	0.946	0.910	0.187	0.140	0.186	0.555	0.849
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.796	0.852	0.827	0.889	0.912	0.735	0.797	0.810	0.850	0.946	0.874	0.921	0.850	0.939	0.885	0.206	0.229	0.247	0.540	0.820
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.748	0.812	0.796	0.882	0.920	0.662	0.707	0.735	0.828	0.936	0.862	0.957	0.874	0.945	0.909	0.089	0.053	0.258	0.441	0.713
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.779	0.812	0.852	0.852	0.912	0.719	0.711	0.807	0.805	0.921	0.856	0.951	0.909	0.908	0.903	0.188	0.059	0.356	0.601	0.738
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.755	0.809	0.775	0.861	0.891	0.659	0.776	0.687	0.801	0.878	0.886	0.849	0.891	0.933	0.908	0.114	0.246	0.210	0.521	0.567
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.778	0.841	0.834	0.888	0.913	0.742	0.785	0.783	0.837	0.896	0.826	0.916	0.905	0.952	0.933	0.254	0.191	0.340	0.501	0.720
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.752	0.848	0.816	0.888	0.916	0.702	0.780	0.754	0.834	0.934	0.814	0.933	0.899	0.952	0.903	0.180	0.182	0.223	0.442	0.808
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.809	0.833	0.837	0.878	0.916	0.759	0.764	0.786	0.836	0.891	0.874	0.922	0.903	0.933	0.946	0.264	0.210	0.269	0.443	0.741
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.810	0.832	0.849	0.859	0.925	0.769	0.766	0.800	0.801	0.901	0.863	0.916	0.916	0.933	0.952	0.175	0.139	0.278	0.560	0.560
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.805	0.844	0.833	0.888	0.922	0.771	0.778	0.786	0.836	0.930	0.850	0.928	0.898	0.951	0.916	0.191	0.172	0.233	0.470	0.839
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.740	0.817	0.802	0.878	0.921	0.660	0.719	0.753	0.834	0.919	0.844	0.951	0.861	0.933	0.926	0.156	0.092	0.186	0.473	0.856
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.742	0.807	0.786	0.863	0.904	0.657	0.695	0.710	0.808	0.906	0.856	0.964	0.885	0.928	0.903	0.060	0.241	0.192	0.519	0.589
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.761	0.790	0.761	0.818	0.820	0.622	0.700	0.640	0.750	0.791	0.982	0.909	0.946	0.903	0.854	0.066	0.064	0.078	0.417	0.083
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.759	0.808	0.804	0.853	0.881	0.689	0.741	0.743	0.792	0.936	0.850	0.891	0.880	0.926	0.836	0.060	0.119	0.167	0.585	0.693
3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.750	0.813	0.803	0.859	0.893	0.673	0.713	0.739	0.807	0.916	0.850	0.952	0.886	0.921	0.879	0.158	0.087	0.229	0.603	0.788
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$	0.754	0.812	0.794	0.878	0.909	0.694	0.754	0.731	0.827	0.863	0.831	0.886	0.874	0.939	0.963	0.115	0.223	0.238	0.440	0.572
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.770	0.833	0.773	0.865	0.912	0.689	0.752	0.711	0.822	0.924	0.874	0.940	0.850	0.915	0.903	0.096	0.258	0.114	0.577	0.808
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.784	0.835	0.841	0.884	0.931	0.730	0.775	0.778	0.839	0.930	0.850	0.910	0.922	0.940	0.934	0.229	0.244	0.222	0.579	0.850

OCC-DNA in the sensor array (4 variables)	F1-score					PPV					Sensitivity					Sensitivity at 98% specificity				
	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.783	0.823	0.798	0.848	0.881	0.752	0.745	0.751	0.781	0.856	0.819	0.922	0.855	0.934	0.909	0.133	0.161	0.169	0.572	0.552
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.776	0.851	0.856	0.880	0.945	0.705	0.767	0.795	0.832	0.952	0.868	0.958	0.934	0.940	0.939	0.194	0.154	0.328	0.574	0.874
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.793	0.824	0.826	0.869	0.905	0.744	0.718	0.762	0.813	0.897	0.856	0.969	0.903	0.939	0.916	0.183	0.064	0.232	0.491	0.819
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$	0.731	0.814	0.787	0.873	0.896	0.673	0.710	0.743	0.825	0.929	0.808	0.958	0.844	0.934	0.872	0.067	0.072	0.341	0.448	0.607
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.802	0.851	0.808	0.882	0.928	0.751	0.783	0.786	0.828	0.930	0.862	0.933	0.838	0.945	0.927	0.175	0.175	0.245	0.508	0.848
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.775	0.846	0.838	0.886	0.923	0.741	0.783	0.783	0.848	0.914	0.826	0.927	0.910	0.933	0.933	0.210	0.177	0.310	0.511	0.704
No 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.775	0.808	0.848	0.848	0.912	0.738	0.754	0.800	0.787	0.923	0.819	0.880	0.910	0.921	0.903	0.158	0.165	0.276	0.585	0.807
No 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.769	0.778	0.764	0.844	0.875	0.641	0.700	0.619	0.781	0.873	0.964	0.879	1.000	0.922	0.885	0.060	0.140	0.090	0.432	0.369
No 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.735	0.806	0.795	0.852	0.880	0.667	0.734	0.719	0.788	0.909	0.820	0.910	0.898	0.933	0.860	0.097	0.130	0.226	0.501	0.676
No 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.727	0.805	0.793	0.848	0.878	0.665	0.720	0.733	0.787	0.924	0.808	0.921	0.874	0.921	0.843	0.060	0.088	0.233	0.457	0.667
No 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.792	0.823	0.812	0.862	0.898	0.731	0.731	0.752	0.812	0.897	0.868	0.945	0.885	0.922	0.903	0.146	0.083	0.266	0.467	0.690
No 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.737	0.800	0.758	0.868	0.916	0.660	0.784	0.666	0.828	0.861	0.838	0.825	0.892	0.915	0.982	0.096	0.161	0.050	0.499	0.319
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$	0.730	0.827	0.810	0.879	0.900	0.661	0.768	0.731	0.836	0.918	0.820	0.904	0.915	0.932	0.890	0.102	0.305	0.205	0.480	0.836
No 4-N(C_2H_5)$_2$*$(TAT)_4$	0.790	0.850	0.814	0.881	0.916	0.735	0.774	0.765	0.845	0.962	0.857	0.946	0.875	0.928	0.879	0.212	0.169	0.242	0.551	0.831
No 3,4,5-F_3*$CT_2C_3T_2C$	0.779	0.837	0.842	0.882	0.905	0.737	0.791	0.800	0.836	0.927	0.832	0.897	0.898	0.939	0.886	0.248	0.078	0.263	0.560	0.756
No 3,4,5-F_3*$(TAT)_4$	0.783	0.824	0.798	0.859	0.914	0.738	0.747	0.736	0.794	0.917	0.837	0.921	0.873	0.940	0.915	0.115	0.119	0.121	0.579	0.771
No 3,4,5-F_3*$(AT)_{15}$	0.731	0.825	0.754	0.876	0.889	0.671	0.816	0.687	0.832	0.840	0.808	0.843	0.844	0.928	0.952	0.054	0.186	0.162	0.464	0.284
No 3,4,5-F_3*$(GT)_{15}$	0.769	0.804	0.780	0.840	0.894	0.674	0.715	0.721	0.769	0.898	0.898	0.928	0.856	0.927	0.896	0.060	0.094	0.155	0.538	0.812
All six OCC-DNA sensors	0.728	0.788	0.750	0.831	0.858	0.667	0.736	0.724	0.780	0.874	0.808	0.855	0.784	0.897	0.848	0.042	0.095	0.180	0.327	0.532
All six OCC-DNA sensors with 24h incubation	0.772	0.815	0.825	0.863	0.908	0.714	0.715	0.749	0.795	0.869	0.844	0.951	0.922	0.945	0.951	0.155	0.184	0.247	0.575	0.318
All six OCC-DNA sensors with 72h incubation	0.748	0.836	0.844	0.894	0.926	0.683	0.755	0.768	0.822	0.928	0.831	0.940	0.939	0.982	0.927	0.031	0.123	0.132	0.595	0.891
All six OCC-DNA sensors with 2 h and 24 h incubation	0.763	0.830	0.833	0.878	0.919	0.742	0.735	0.797	0.812	0.896	0.802	0.958	0.880	0.958	0.945	0.197	0.079	0.419	0.754	0.813
All six OCC-DNA sensors with 2 h, 24 h, and 72 h incubation	0.867	0.848	0.877	0.892	0.927	0.816	0.781	0.827	0.842	0.963	0.928	0.934	0.939	0.951	0.897	0.141	0.113	0.308	0.715	0.868

(3) Six spectra variables, *dwl + dwl* + dint + dint* + Δwl + Δint (N_f = 6):*

OCC-DNA in the sensor array (6 variables)	F1-score					PPV					Sensitivity					Sensitivity at 98% specificity				
	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$	0.735	0.796	0.801	0.865	0.894	0.735	0.727	0.748	0.770	0.836	0.877	0.753	0.860	0.842	0.903	0.915	0.246	0.197	0.276	0.615
4-N(C_2H_5)$_2$*$(TAT)_4$	0.731	0.807	0.825	0.843	0.887	0.731	0.676	0.701	0.778	0.786	0.883	0.802	0.958	0.885	0.915	0.896	0.175	0.166	0.261	0.522
3,4,5-F_3*$CT_2C_3T_2C$	0.738	0.827	0.753	0.836	0.909	0.738	0.699	0.736	0.625	0.796	0.939	0.790	0.947	0.952	0.884	0.885	0.191	0.113	0.147	0.529
3,4,5-F_3*$(TAT)_4$	0.762	0.843	0.829	0.889	0.921	0.762	0.742	0.786	0.762	0.848	0.898	0.789	0.916	0.917	0.940	0.946	0.288	0.159	0.258	0.599
3,4,5-F_3*$(AT)_{15}$	0.742	0.800	0.837	0.865	0.914	0.742	0.718	0.722	0.775	0.795	0.919	0.771	0.903	0.915	0.951	0.915	0.125	0.059	0.191	0.566
3,4,5-F_3*$(GT)_{15}$	0.738	0.833	0.814	0.870	0.899	0.738	0.730	0.768	0.776	0.821	0.904	0.752	0.917	0.868	0.928	0.896	0.209	0.130	0.255	0.619
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$	0.770	0.796	0.858	0.877	0.918	0.770	0.746	0.723	0.810	0.819	0.867	0.801	0.891	0.916	0.946	0.976	0.192	0.126	0.290	0.584
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.751	0.822	0.829	0.882	0.919	0.751	0.736	0.786	0.799	0.845	0.923	0.771	0.866	0.874	0.928	0.916	0.294	0.095	0.318	0.603
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.729	0.771	0.745	0.834	0.902	0.729	0.673	0.732	0.675	0.800	0.885	0.801	0.831	0.837	0.878	0.921	0.144	0.138	0.078	0.470
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.749	0.834	0.854	0.889	0.927	0.749	0.748	0.771	0.827	0.841	0.936	0.760	0.915	0.892	0.945	0.921	0.203	0.199	0.242	0.659
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.731	0.775	0.742	0.813	0.853	0.731	0.671	0.711	0.678	0.779	0.806	0.808	0.867	0.824	0.854	0.909	0.049	0.062	0.208	0.400
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.778	0.811	0.858	0.883	0.910	0.778	0.744	0.714	0.807	0.822	0.910	0.820	0.940	0.928	0.958	0.915	0.150	0.108	0.232	0.579
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$	0.739	0.811	0.851	0.883	0.920	0.739	0.690	0.727	0.795	0.826	0.934	0.801	0.927	0.922	0.951	0.909	0.279	0.144	0.206	0.568
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.723	0.810	0.790	0.858	0.914	0.723	0.683	0.742	0.722	0.806	0.880	0.771	0.896	0.874	0.926	0.952	0.167	0.093	0.255	0.594
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.736	0.795	0.791	0.863	0.898	0.736	0.681	0.737	0.727	0.824	0.919	0.808	0.875	0.874	0.909	0.879	0.222	0.233	0.227	0.637
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.729	0.820	0.779	0.893	0.901	0.729	0.668	0.784	0.760	0.853	0.852	0.808	0.867	0.814	0.940	0.957	0.167	0.114	0.271	0.685
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.770	0.807	0.846	0.871	0.917	0.770	0.739	0.717	0.795	0.800	0.939	0.807	0.928	0.910	0.958	0.898	0.211	0.094	0.181	0.512
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.783	0.831	0.825	0.862	0.890	0.783	0.743	0.751	0.756	0.799	0.842	0.832	0.935	0.922	0.940	0.946	0.178	0.171	0.176	0.579
3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.735	0.809	0.810	0.852	0.862	0.735	0.678	0.748	0.730	0.795	0.853	0.808	0.885	0.916	0.921	0.879	0.103	0.153	0.186	0.537
3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.729	0.846	0.761	0.872	0.892	0.729	0.668	0.821	0.750	0.835	0.858	0.808	0.875	0.777	0.916	0.933	0.031	0.221	0.315	0.611
3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.730	0.822	0.754	0.843	0.846	0.730	0.667	0.743	0.707	0.785	0.779	0.814	0.929	0.824	0.916	0.928	0.060	0.167	0.144	0.401
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.808	0.826	0.847	0.875	0.916	0.808	0.763	0.720	0.790	0.817	0.932	0.862	0.970	0.921	0.945	0.903	0.204	0.076	0.200	0.542
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$	0.826	0.846	0.867	0.890	0.923	0.826	0.776	0.764	0.823	0.824	0.937	0.886	0.951	0.922	0.969	0.914	0.286	0.192	0.310	0.492
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.819	0.822	0.868	0.876	0.938	0.776	0.736	0.831	0.811	0.921	0.874	0.933	0.915	0.957	0.957	0.221	0.090	0.286	0.591	0.843
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.783	0.786	0.813	0.868	0.913	0.749	0.677	0.767	0.817	0.926	0.825	0.940	0.873	0.932	0.903	0.198	0.061	0.342	0.576	0.817

OCC-DNA in the sensor array (6 variables)	F1-score					PPV					Sensitivity					Sensitivity at 98% specificity				
	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.735	0.796	0.816	0.867	0.916	0.678	0.723	0.778	0.812	0.927	0.808	0.896	0.867	0.933	0.908	0.176	0.158	0.287	0.564	0.662
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.794	0.845	0.817	0.876	0.921	0.766	0.800	0.775	0.814	0.923	0.825	0.903	0.867	0.951	0.921	0.311	0.232	0.352	0.660	0.750
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.836	0.831	0.855	0.881	0.936	0.800	0.742	0.819	0.815	0.941	0.881	0.946	0.899	0.964	0.933	0.299	0.066	0.299	0.685	0.825
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.774	0.795	0.816	0.880	0.914	0.721	0.736	0.782	0.840	0.928	0.843	0.873	0.861	0.927	0.903	0.260	0.141	0.219	0.503	0.664
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.812	0.804	0.876	0.863	0.926	0.780	0.686	0.835	0.808	0.934	0.850	0.976	0.928	0.927	0.921	0.257	0.036	0.208	0.688	0.744
4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.741	0.818	0.807	0.865	0.912	0.699	0.781	0.774	0.808	0.885	0.795	0.868	0.856	0.933	0.945	0.170	0.163	0.303	0.636	0.450
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.786	0.815	0.878	0.887	0.935	0.759	0.764	0.838	0.844	0.927	0.826	0.885	0.935	0.939	0.946	0.257	0.182	0.407	0.644	0.832
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.794	0.861	0.849	0.887	0.921	0.757	0.796	0.791	0.835	0.921	0.837	0.946	0.922	0.951	0.927	0.299	0.189	0.274	0.623	0.813
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.807	0.835	0.869	0.890	0.944	0.771	0.753	0.831	0.836	0.933	0.856	0.946	0.916	0.957	0.957	0.275	0.150	0.367	0.581	0.827
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.787	0.796	0.858	0.874	0.927	0.766	0.718	0.812	0.819	0.905	0.814	0.897	0.915	0.939	0.951	0.155	0.090	0.288	0.619	0.809
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.819	0.827	0.883	0.884	0.945	0.795	0.765	0.834	0.820	0.947	0.850	0.910	0.946	0.964	0.945	0.288	0.177	0.281	0.569	0.879
4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.768	0.804	0.823	0.879	0.905	0.733	0.718	0.786	0.827	0.942	0.812	0.921	0.868	0.939	0.873	0.303	0.182	0.347	0.626	0.860
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.735	0.802	0.795	0.856	0.900	0.679	0.723	0.735	0.796	0.934	0.808	0.911	0.880	0.933	0.872	0.133	0.185	0.263	0.584	0.686
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.758	0.771	0.778	0.834	0.832	0.632	0.715	0.717	0.782	0.819	0.952	0.857	0.855	0.898	0.849	0.060	0.153	0.238	0.489	0.400
3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.745	0.819	0.807	0.861	0.894	0.708	0.757	0.770	0.800	0.914	0.789	0.897	0.856	0.933	0.879	0.263	0.070	0.227	0.496	0.524
3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.738	0.824	0.814	0.872	0.916	0.705	0.743	0.774	0.813	0.890	0.782	0.933	0.867	0.944	0.945	0.311	0.203	0.300	0.606	0.825
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 4-N(C_2H_5)$_2$*$(TAT)_4$	0.742	0.802	0.829	0.870	0.913	0.697	0.733	0.782	0.817	0.939	0.800	0.891	0.892	0.933	0.891	0.279	0.196	0.345	0.655	0.778
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.768	0.822	0.834	0.889	0.919	0.716	0.750	0.794	0.845	0.939	0.836	0.915	0.886	0.939	0.903	0.295	0.165	0.269	0.687	0.795
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.822	0.818	0.846	0.872	0.930	0.784	0.759	0.796	0.812	0.934	0.868	0.891	0.910	0.945	0.927	0.293	0.242	0.360	0.589	0.749

OCC-DNA in the sensor array (6 variables)	F1-score					PPV					Sensitivity					Sensitivity at 98% specificity				
	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM	Logi	DT	ANN	RF	SVM
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.778	0.824	0.817	0.850	0.895	0.747	0.742	0.777	0.787	0.917	0.818	0.933	0.873	0.927	0.879	0.257	0.146	0.217	0.578	0.746
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.811	0.841	0.886	0.883	0.935	0.774	0.821	0.842	0.836	0.915	0.857	0.873	0.940	0.939	0.957	0.269	0.396	0.324	0.661	0.797
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$CT_2C_3T_2C$	0.785	0.831	0.866	0.880	0.925	0.753	0.724	0.797	0.820	0.952	0.825	0.976	0.952	0.951	0.903	0.241	0.120	0.258	0.626	0.837
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(TAT)_4$	0.734	0.804	0.818	0.861	0.912	0.693	0.786	0.780	0.810	0.913	0.789	0.831	0.874	0.921	0.915	0.246	0.151	0.352	0.552	0.723
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.796	0.839	0.863	0.887	0.931	0.750	0.769	0.804	0.833	0.944	0.850	0.928	0.934	0.951	0.921	0.310	0.102	0.399	0.670	0.774
No 4-N(C_2H_5)$_2$*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.782	0.821	0.875	0.888	0.933	0.770	0.780	0.820	0.829	0.906	0.802	0.874	0.946	0.957	0.964	0.281	0.109	0.360	0.664	0.726
No 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(TAT)_4$	0.786	0.822	0.848	0.868	0.924	0.758	0.715	0.803	0.804	0.924	0.819	0.970	0.904	0.945	0.927	0.202	0.059	0.171	0.646	0.763
No 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(AT)_{15}$	0.764	0.775	0.780	0.841	0.875	0.623	0.708	0.715	0.771	0.824	0.988	0.872	0.861	0.926	0.939	0.078	0.111	0.144	0.500	0.103
No 3,4,5-F_3*$CT_2C_3T_2C$ + 3,4,5-F_3*$(GT)_{15}$	0.730	0.809	0.818	0.871	0.899	0.696	0.757	0.758	0.819	0.904	0.771	0.891	0.892	0.933	0.902	0.229	0.211	0.119	0.586	0.672
No 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(AT)_{15}$	0.738	0.794	0.804	0.863	0.891	0.701	0.741	0.735	0.811	0.840	0.783	0.860	0.891	0.927	0.951	0.259	0.240	0.174	0.513	0.229
No 3,4,5-F_3*$(TAT)_4$ + 3,4,5-F_3*$(GT)_{15}$	0.801	0.826	0.834	0.882	0.894	0.758	0.720	0.773	0.830	0.921	0.855	0.970	0.909	0.945	0.873	0.223	0.076	0.295	0.547	0.813
No 3,4,5-F_3*$(AT)_{15}$ + 3,4,5-F_3*$(GT)_{15}$	0.751	0.797	0.808	0.860	0.893	0.709	0.749	0.764	0.829	0.922	0.813	0.860	0.861	0.897	0.872	0.208	0.061	0.374	0.548	0.640
No 4-N(C_2H_5)$_2$*$CT_2C_3T_2C$	0.741	0.808	0.816	0.877	0.916	0.691	0.728	0.774	0.830	0.914	0.808	0.914	0.881	0.933	0.920	0.194	0.142	0.346	0.641	0.632
No 4-N(C_2H_5)$_2$*$(TAT)_4$	0.816	0.842	0.840	0.886	0.931	0.772	0.762	0.787	0.814	0.925	0.868	0.946	0.904	0.976	0.939	0.329	0.146	0.401	0.599	0.885
No 3,4,5-F_3*$CT_2C_3T_2C$	0.782	0.824	0.869	0.873	0.933	0.759	0.760	0.850	0.814	0.905	0.813	0.904	0.898	0.945	0.964	0.270	0.119	0.349	0.680	0.818
No 3,4,5-F_3*$(TAT)_4$	0.783	0.824	0.841	0.866	0.924	0.738	0.735	0.788	0.820	0.951	0.836	0.940	0.904	0.920	0.902	0.246	0.151	0.330	0.662	0.824
No 3,4,5-F_3*$(AT)_{15}$	0.734	0.795	0.765	0.861	0.896	0.678	0.708	0.737	0.802	0.843	0.808	0.916	0.801	0.933	0.958	0.150	0.085	0.244	0.610	0.131
No 3,4,5-F_3*$(GT)_{15}$	0.742	0.806	0.804	0.855	0.897	0.690	0.719	0.758	0.786	0.880	0.807	0.922	0.862	0.939	0.921	0.276	0.052	0.168	0.488	0.532
All six OCC-DNA sensors	0.728	0.792	0.729	0.856	0.874	0.667	0.730	0.668	0.811	0.847	0.808	0.878	0.808	0.909	0.910	0.059	0.092	0.048	0.448	0.201
All six OCC-DNA sensors with 24h incubation	0.781	0.818	0.852	0.851	0.902	0.734	0.734	0.785	0.793	0.851	0.837	0.928	0.933	0.920	0.963	0.142	0.142	0.246	0.541	0.231
All six OCC-DNA sensors with 72h incubation	0.777	0.825	0.886	0.890	0.930	0.722	0.771	0.832	0.825	0.931	0.849	0.892	0.951	0.969	0.933	0.036	0.086	0.344	0.708	0.880
All six OCC-DNA sensors with 2 h and 24 h incubation	0.839	0.821	0.885	0.864	0.936	0.786	0.724	0.829	0.793	0.968	0.905	0.951	0.952	0.951	0.908	0.232	0.144	0.369	0.725	0.867
All six OCC-DNA sensors with 2 h, 24 h, and 72 h incubation	0.872	0.831	0.900	0.888	0.940	0.818	0.764	0.847	0.832	0.939	0.940	0.915	0.964	0.957	0.946	0.278	0.142	0.324	0.675	0.899

Rebuttal 2

Reviewer #1:

In this revised manuscript, the authors have made their effort to address the issues and comments raised by the reviewers. The quality of this manuscript has been improved after the substantial revisions and appended discussions on the design rationale, superiorities over the conventional tests, and bias analysis. Although the underlying interactions and types of proteins that contribute to the signal changes of the sensors remain unclear, it may be investigated using proteome analysis in the future to benefit pathophysiology study. However, the following minor concerns need to be addressed before the acceptance of manuscript.

We thank the reviewer for their constructive assessment of the manuscript and our responses. We have addressed the additional comments below.

1. The authors mentioned that the proposed sensors have non-stringent storage requirements and low screening cost compared to CA125 blood test. Then, what is the shelf life of these DNA-wrapped SWCNTs? What are the required storage conditions of these SWCNTs? If stored at ambient conditions, how long can DNA coating maintain its integrity and produce reproducible results? The authors should provide data on this. I think that the authors should tone down on the cost statement.

We appreciate the reviewer's comment on the storage and shelf-life. The OCC-DNAs have thus far been stored at 4 ^{o}C until used. To assess long-term stability, we compared the absorbance spectra using OCC-DNA samples that were stored for 6-10 months at 4 ^{o}C after synthesis (see Figure below). We observed no significant decrease, and overall minor changes, in the absorbance bands, indicating that the OCC-DNAs remained colloidally stable. We included the figure as **Fig S12** (and pasted below) and added the experimental details in the Method section. For the data collection of the training set, we used OCC-DNAs that were synthesized within 6 months prior to testing with patient serum samples. For the test set, we used freshly prepared OCC-DNAs (less than 2 weeks old). Our response data suggest that the shelf-life of the OCC-DNAs is at least 6 months at 4^{o}C. We note that storage at this temperature is easier than the cold chain logistics needed for the storage of antibodies (normally -20^{o}C). Although it is known that DNA-wrapped SWCNTs can be stored at ambient temperatures for extended durations, we did not yet investigate this possibility for OCC-DNAs. We modified/moderated the discussion of experimental conditions and costs in the manuscript text accordingly.

Large scale synthesis of OCC-DNAs in Methods
"The OCC-SWCNTs were stabilized by 3.5 mg/mL ssDNA in phosphate buffered saline (PBS). The OCC-SWCNT were individually dispersed by ultrasonication at 6W for 60 min using a probe-tip sonicator (Sonics & Materials, Inc) at 4^{o}C for 1 hour. The DNA to SWCNT mass ratio is 5 to 1. Then the OCC-DNA solutions were centrifuged at 100,000 g and 4 ^{o}C for 30 min. The 80% supernatant was dialyzed against PBS for 36 hours to remove free DNA (Spectra-Por, Float-A-Lyzer, MWCO = 1MDa). The absorption spectra of the dialyzed solutions were collected with a UV-Vis-NIR spectrophotometer (Jasco, Tokyo, Japan). After subtracting background, the optical density at (6,5) E_{11} (~1000 nm) was used to estimate the relative OCC-DNA concentration (Fig S11). The OCC-DNAs were kept at 4^{o}C until used (up to 6 months) as the OCC-DNAs remained colloidally stable (**Fig S12**)."

OCC-DNA and serum/recombinant protein handling in Methods
"For the training set data collection, we used the OCC-DNAs that were synthesized within 6 months prior to testing with patient serum samples (1 week to 6 months old). For the test set, we used freshly prepared OCC-DNAs (less than 2 weeks old). The OCC-DNA concentration was adjusted to 0.325 mg/L in PBS. We introduced 20 µL of a patient serum sample to 80 µL of OCC-DNAs in a 96-well plate (Corning) to make the OCC-DNA concentration of 0.26 mg/L in each well. OCC-DNAs in 100 µL PBS (0.26 mg/L) was also prepared to compare the relative changes in sensor response in serum for feature vector construction (See Data preprocessing in Methods). The OCC-DNA was incubated at room temperature for 2 hours and in a cold room (4 ^{o}C) after the spectral acquisition at 2-hour time point. Data were taken at three time points during incubation: 2 hours, 24 hours, and 72 hours."

Page 9,
“Fourth, because the technology does not rely on antibody-based molecular recognition elements, the sensors could be more robust than existing methods, enabling use in resource-limited settings and in technologies such as point-of-care and wearable/implantable devices (Lee *et al ACS Sens.* **2019**). Lastly, the sensor technology also has the potential to be developed as a relatively inexpensive and rapid screening tool to result in a single, easy-to-interpret test result in primary care settings. The materials needed for the sensor cost approximately $5 per sample because of the small amount of OCC-DNAs needed for screening (<5 nanograms). The cost of the sensor measurement would also diminish if measured via high-throughput instruments, and the potential for the use of very-low sample volumes is substantial.”

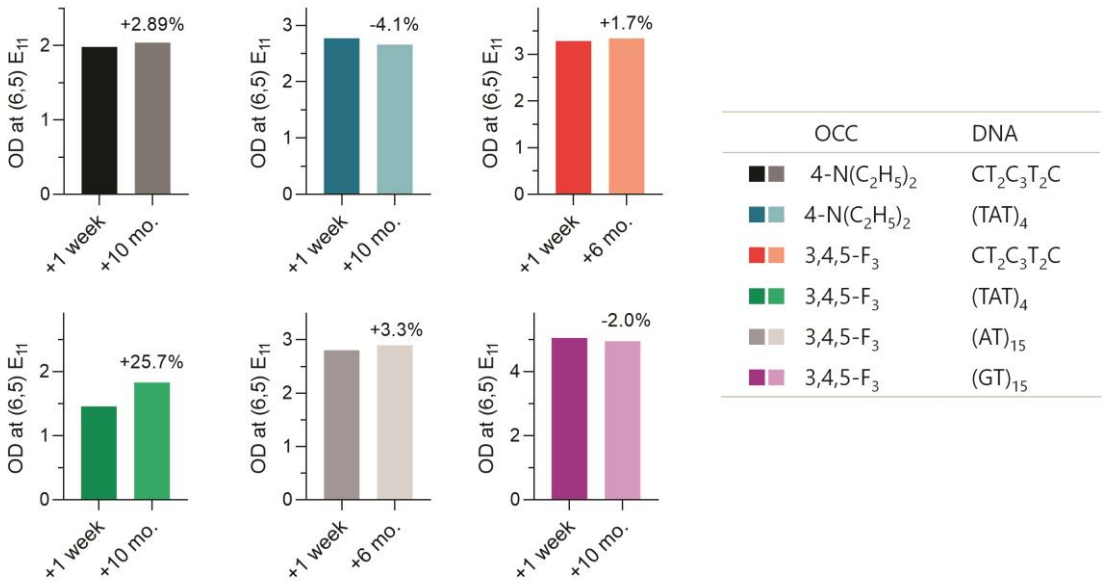

Fig S12 | Long term stability of OCC-DNAs. Optical density of (6,5) E_{11} bands of the OCC-DNAs, measured 1 week and 6-10 months after synthesis. The OCC-DNAs in 1XPBS were stored at 4 ^{o}C until measured.

2. How is the reproducibility of the properties and test results of OCC-DNAs from different batches? It is known that wrapped SWCNT might differ in bioaffinity from batch to batch.

We thank the reviewer’s question regarding the reproducibility of the test results of OCC-DNAs from different batches. To validate the reproducibility and generalizability of our sensor technology, the test set data was collected using a different batch of OCC-DNAs from that of the training set. The new batch of OCC-DNAs was synthesized under the same conditions as the first batch but several months after the first. We added new discussion on the reproducibility of the OCC-DNAs and results as follows:

Page 5
“To further assess the robustness of the sensor array and algorithm, we synthesized a new batch of OCC-DNAs under the same condition and collected the sensor array response data to an independent test set of 54 patient samples (N_{sa} = 54). To evaluate the model performance in various medical conditions, the test set was sampled from different patients, comprised of 7 HGSOC, 5 other gynecologic diseases, 32 non-gynecologic diseases, and 10 heathy patients. With this new sample set, the optimized SVM model resulted in 100% sensitivity at 98% specificity and an F-score of 0.978. These values are consistent with the cross-validation scores and gave a similar receiver operating characteristic (ROC) curve (**Fig 3d**), indicating that the model did not overfit the data.”

3. How good is the correlation of the test results with the staging of ovarian cancer?

The optimized prediction models identified both stage III and IV cancers but virtually no earlier-stage samples were available, due to the rarity of early-stage detection via existing methods. We did not observe statistically significant differences in the accuracy between stage III and IV cancers.

4. What is the earliest cancer stage that the proposed in your sample population?

We appreciate the reviewer's comment. Due to the difficulty in diagnosing early-stage ovarian cancer, sufficient numbers of early-stage samples were not available. Over 95% of samples were from stage III and IV patients in the training set. The optimized SVM model identified the one stage II HGSOC sample in the test set. To better discuss the clinical potential of our sensor technology, we have modified the text accordingly:

Page 8:
"Model performance of the sensor technology exceeds the results of current best clinical screening test using longitudinal CA125 and second-line transvaginal ultrasonography (Jacobs *et al Lancet* **2015**) (87% vs. 84% clinical sensitivities at 98% specificity). However, considering the fact that specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, prediction outcomes may differ in clinical screening setting in which specimens are obtained in asymptomatic individuals before clinical diagnosis. Further studies in a high-risk cohort, such as BRCA mutation carriers undergoing risk-reducing surgery, are warranted to evaluate the ability of the technology to identify pre-invasive and early-invasive disease."

Reviewer #2

Reviewer 2 notes the improvements the authors made to the manuscript and suggest for publication in Nature Biomedical Engineering.

We appreciate the reviewer's assistance in enabling these improvements to the manuscript.

Reviewer #3

Major issues

We thank the reviewer's support to enable the improvements in the clinical significance of the manuscript.

1. The authors use an estimate of 87% sensitivity at 98% specificity. However, given the results in Fig. 3, the more appropriate and robust estimate is 77% sensitivity at 98% specificity, achieved by weighting the PPV more than the sensitivity (weight ß = 0.2). The authors do make a caveat correctly that their markers achieve this sensitivity in clinically diagnosed patients, while the screening studies (UKCTOCS) are in asymptomatic patients where sensitivity is expected to be significantly lower.

We understand this concern regarding Figure 3 and its conclusions. We realize that its presentation led to confusion, which prompted us to modify the data presentation and description in the main text. Briefly, each data point in Fig 3d, at each β value, indicated the sensitivity of the 10-fold cross-validation of a different nanosensor array combination. Therefore, the single highest data point in the figure denoted the best prediction model, indicating is the sensitivity of that sensor array. The best performing model that yielded 87% sensitivity at 98% specificity was the sensor array combination of 4-N$(C_2H_5)_2$*$CT_2C_3T_2C$, 4-N$(C_2H_5)_2$*$(TAT)_4$, 3,4,5-F_3*$(TAT)_4$, 3,4,5-F_3*$(AT)_{15}$, and 3,4,5-F_3*$(GT)_{15}$, optimized with β=1, highlighted as red dots in the revised **Fig 3d**. The median sensitivity (gray line) is simply a guide to the eye for the reader, to emphasize the importance of using a weighted harmonic mean in the model development process to maximize sensitivity at 98% specificity. Thus, the claim regarding the results of Fig 3d, that the prediction model of the sensor array yields a sensitivity at 98%

specificity of as high as 87%, is valid. The caption and presentation of **Fig 3d** is now updated to clarify the data interpretation.

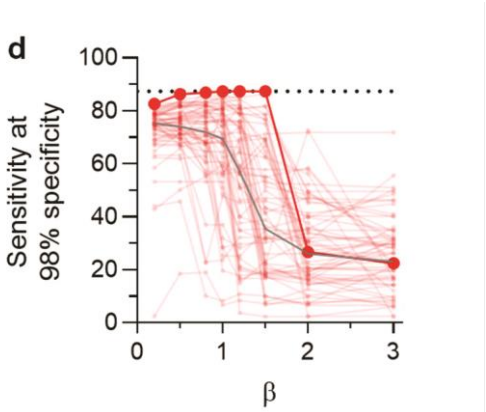

Fig. 3 | Optimization of machine learning algorithms for HGSOC classification. … d, Sensitivities, of all possible sensor array combinations, composed of up to 6 OCC-DNAs, at 98% specificity, as a function of β in F_β scoring via SVM. Small red dots denote sensitivities of each nanosensor array at varying β. Large red dots denote the overall best-performing nanosensor combination. The gray line is the median of sensitivities for all optimized nanosensor arrays.

2. In response to my recommendation (#2) to maximize sensitivity at 98% specificity by using a weighted harmonic mean. They argue that weights less than 1.0 weigh false positives more importantly than true positives. They investigated weights from 3 down to 0.2, with the smaller weights systematically increasing sensitivity at 98% specificity, although this increase was statistically insignificant. However, the median sensitivity changes from a low of 25% to a high of 77%. This change is of great clinical significance, and it's not clear what test was used to conclude the change was not statistically significant. I am not clear why the authors conclude that their method yields a sensitivity of 87% at 98% specificity when ß is 1.0 since the graph (Fig. 3 under "Page 5") clearly shows the optimal sensitivity is 77% when ß is 0.2. The authors continue to claim 87% sensitivity at 98% specificity. Fig. 3 shows that 77% is the more appropriate estimate of sensitivity at 98% specificity, and this sensitivity estimate should be the comparison with the UKCTOCS results of 84% sensitivity at 98% specificity.
The issue of specificity for ovarian cancer screening has been addressed and overcome by Jacobs' series of screening studies showing that a multimodal screen with a blood test (CA125) followed by TVS for patients with a positive blood test yields an acceptable PPV and thus is sufficiently specific for screening. This series of studies stands in contrast to the authors' assertion in the Abstract that biomarkers are not specific for ovarian cancer. The challenge for ovarian cancer biomarkers as a first line test leading to imaging as a second line test is not specificity, it is that the initial test is not sensitive enough to detect a high proportion of cases in early stage. This challenge of low sensitivity for early-stage disease should be the clear statement of the clinical challenge facing biomarker researchers in identifying a first line test for early detection of ovarian cancer.

We agree with the reviewer that optimizing the prediction model weighing PPV more than sensitivity (lower β value) can improve clinical significance, and we have now changed the manuscript to better address this aspect. In the optimization using a weighted harmonic mean, the median sensitivity gradually improved from 26% to 75% when β decreases from 3 to 0.2 (gray line in Fig 3d). To quantitatively assess whether different β value resulted in statistically significant improvement in sensitivity at 98% specificity, we performed the Holm-Šídák's multiple comparisons after ANOVA (the summary of tests is included in the end of the rebuttal). The analysis compares the mean of each group with the mean of every other group. Here, each group comprises sensitivity of 10-fold cross-validation of all the possible sensor array combinations (63 in total) at different weight. We observed a significant improvement in the sensitivity at 98% specificity at decreasing β from 3 to 1. On the other hand, the adjusted p-values in the comparisons β=1 vs. β=0.2, 0.5, and 0.8 were larger than 0.1, indicating that weighing more in PPV (1 > β) resulted in statistically insignificant difference. We now included the multiple comparison test as **Table S4** and modified the main text to clarify these analyses.

Page 5:
"At decreasing β from 3 to 0.2, sensitivity at 98% specificity systematically increased although the improvement was statistically insignificant between β values of 0.2, 0.5, 0.8, and 1 in this study (**Table S4**). The best performing prediction model was the sensor array combination of 4-$N(C_2H_5)_2$*$CT_2C_3T_2C$, 4-$N(C_2H_5)_2$*$(TAT)_4$, 3,4,5-F_3*$(TAT)_4$, 3,4,5-F_3*$(AT)_{15}$, and 3,4,5-F_3*$(GT)_{15}$, and yielded 87% sensitivity at 98% specificity when PPV and sensitivity were equally weighted (β =1)."

Regarding the correct estimate of sensitivity at 98% specificity, as explained in the response to Comment #1, we believe that the best performing model with 87% sensitivity at 98% specificity should be compared with the UKCTOCS results. Each data point in Fig 3d, at each β value, indicated the sensitivity of the 10-fold cross-validation of a certain nanosensor array combination. The best prediction model is, therefore, the single highest data point in Fig 3d rather than the highest median sensitivity of all the possible sensor array combinations at certain β value.

Regarding the challenge in ovarian cancer diagnosis, we are grateful for the reviewer's insight into the field. We agree with the reviewer that the challenge in ovarian cancer diagnosis is the lack of a robust first line test that is highly sensitive at high specificity, and we corrected our statement in the abstract. We reasoned that we should not go too deeply into these issues in the abstract; thus we made a more general statement:

In Abstract,
"Serum biomarkers are widely used as diagnostic indicators, but many are not sensitive enough for screening purposes. Ovarian cancer, for example, is challenging to diagnose, in part because the few established biomarkers for the disease do not provide sufficient diagnostic power."

3. The authors assert in the Abstract that the markers lack specificity because of the elevation in other conditions. This is also incorrect in that even though CA125 is elevated in other conditions, CA125 does not lack specificity when followed with TVS for positive CA125 tests. The elevation in these other conditions is sufficiently rare that a high specificity can be achieved despite their (infrequent) presence. In the vast majority of cases where CA125 is elevated when screening, no other conditions are identified, and subsequently the CA125 usually returns to baseline or stops increasing and becomes stable over time. In Page 1, the authors state that "screening strategies are associated with a high rate of false-positive results and a risk of harm from invasive testing". For UKCTOCS, there was not a high rate of false positive results. The literature had set an upper limit of 10 operations to identify 1 ovarian cancer, and UKCTOCS multimodal arm achieved a rate of 5 operations to identify 1 ovarian cancer, half the rate stated in the literature and therefore within acceptable limits (PPV > 10%). The authors should emphasize that the challenge is increasing the sensitivity for detection of early-stage disease while maintaining the already achieved high specificity (98% for blood test, 99.8% for blood test followed by TVS) and high PPV > 20%.

We agree with the reviewer that longitudinal CA125 measurements with ultrasonography can be specific enough to diagnose HGSOC. We also understand the reviewer's point that the challenge is increasing the sensitivity for detection of early-stage disease while maintaining the already achieved high specificity. We corrected the current diagnostic challenge in ovarian cancer in the abstract (discussed re. Comment #2 above) and the introduction as follows:

In Page 1,
"Conventionally, serum biomarker measurements, such as cancer antigen 125 (CA125), are used as the first line test and/or to monitor high-risk women for ovarian cancer. Other complementary serum biomarkers such as human epidermis protein 4 (HE4), chitinase-3-like protein 1 (YKL40), and mesothelin, or panels of biomarkers, have been reported to result in better discriminatory power over CA125-based screening. However, stand-alone biomarker measurements have proven of little mortality benefit due to limited specificity and low positive predictive value (PPV). Longitudinal CA125 measurements combined with transvaginal ultrasonography (TVS), result in improved PPV for ovarian cancer detection, but the benefit of screening is still elusive. Currently, no screening strategy can identify disease at an early enough stage to reduce mortality.

Major opportunities for improving patient outcomes from ovarian cancer include increasing the sensitivity of early-stage detection while maintaining high specificity, triaging patients presenting with adnexal mass, and the detection of minimum-residual disease in treated patients. Accurate detection of known analytes does not always confer high sensitivity and specificity for disease, however. As established serum biomarkers do not sufficiently represent the ovarian cancer disease state, nor are they sensitive enough to achieve precise, early diagnosis to benefit survival rates, they only provide incremental value for improving treatment options, and they often do not reduce costs for patients.

To seek an alternative approach to overcome diagnostic challenges, we investigated a perception-based strategy. …"

4. The new markers' sensitivity estimate of 87% at 98% specificity should be 77% by Fig. 3.

We believe that 87% sensitivity at 98% specificity is the right estimate as explained above (response to Comments #1 and #2).

5. The authors change the clinical challenge from early detection of ovarian cancer in the asymptomatic population to differentiating ovarian cancer from patients with other diseases. They should be consistent. A better approach would be to argue that they are developing the markers to differentiate ovarian cancer from other diseases as a first step, and that in future studies, they will examine the performance of their markers in the context of screening, where samples from healthy individuals will form the bulk of the control group.

We thank the reviewer for this helpful suggestion. As per this comment and #3 above, we modified discussion to improve consistency of the rationale throughout the manuscript. To clarify the discussion of the outlook on page 8, we now note that specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology. In future studies, the sensor technology can be further examined in the context of screening. Additional studies, where samples from healthy individuals will form the bulk of the control group, are warranted to evaluate the ability of the technology to identify pre-invasive and early-invasive disease. We updated the discussion in the manuscript text as follows:

Page 8:
"Model performance of the sensor technology exceeds the results of current best clinical screening test using longitudinal CA125 and second-line transvaginal ultrasonography (Jacobs *et al Lancet* **2015**) (87% vs. 84% clinical sensitivities at 98% specificity). However, considering the fact that specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, prediction outcomes may differ in the clinical screening setting in which specimens are obtained in asymptomatic individuals before clinical diagnosis. Further studies, in the context of screening the general population, are warranted to evaluate the ability of the technology to identify pre-invasive and early-invasive disease."

Table S4 | Holm-Šídák's multiple comparison t-tests of Fig 3d. Effect of weighted harmonic means of PPV and sensitivity in the model optimization process to improve the sensitivity at 98% specificity. The weighing factor, β denotes the importance of sensitivity with respect to PPV in F-scoring. The analysis compares the statistical significance of the improvement in sensitivity of the 10-fold cross-validation of nanosensor arrays, comprised of up to 6 OCC-DNAs, at varying β.

Comparing groups (different β)	Mean Diff.	Below threshold?	Summary	Adjusted P Value
0.2 vs. 0.5	-0.7878	No	**ns**	**0.8057**
0.2 vs. 0.8	2.981	No	**ns**	**0.7283**
0.2 vs. 1	7.428	No	**ns**	**0.1182**
0.2 vs. 1.2	17.33	Yes	****	<0.0001
0.2 vs. 1.5	30.02	Yes	****	<0.0001
0.2 vs. 2	42.95	Yes	****	<0.0001
0.2 vs. 3	45.93	Yes	****	<0.0001
0.5 vs. 0.8	3.769	No	**ns**	**0.6659**
0.5 vs. 1	8.215	No	**ns**	**0.0717**
0.5 vs. 1.2	18.12	Yes	****	<0.0001
0.5 vs. 1.5	30.81	Yes	****	<0.0001
0.5 vs. 2	43.74	Yes	****	<0.0001
0.5 vs. 3	46.71	Yes	****	<0.0001
0.8 vs. 1	4.447	No	**ns**	**0.5952**
0.8 vs. 1.2	14.35	Yes	***	0.0001
0.8 vs. 1.5	27.04	Yes	****	<0.0001
0.8 vs. 2	39.97	Yes	****	<0.0001
0.8 vs. 3	42.95	Yes	****	<0.0001
1 vs. 1.2	9.906	Yes	*	0.0166
1 vs. 1.5	22.59	Yes	****	<0.0001
1 vs. 2	35.52	Yes	****	<0.0001
1 vs. 3	38.50	Yes	****	<0.0001
1.2 vs. 1.5	12.69	Yes	***	0.0008
1.2 vs. 2	25.61	Yes	****	<0.0001
1.2 vs. 3	28.59	Yes	****	<0.0001
1.5 vs. 2	12.93	Yes	***	0.0006
1.5 vs. 3	15.91	Yes	****	<0.0001
2 vs. 3	2.978	No	ns	0.7283

Rebuttal 3

#1. The authors clarify that their estimate of 87% sensitivity at 98% specificity is the maximum performance over “all possible combinations of sensor array combinations, composed of up to 6 OCC-DNAs at 98% specificity”. They helpfully provide the red line in Fig 3d which shows the sensitivity for the optimal combination. Because it is the optimal combination chosen from many combinations, this level of performance is unlikely to validate due to some component of the performance resulting from random chance variation even in the context of cross validation. How much is due to chance is difficult to evaluate in these high-dimensional situations, but due to this effect of random variation the estimate of 87% sensitivity is optimistically upwardly biased, and due to the apparent large number of combinations this bias can be substantial. Because of this intractable mathematical challenge of reducing the estimate by the appropriate amount, the best approach to obtaining an unbiased estimate of performance is to validate the optimal marker combination in an independent sample set of cases and controls, often from different sources than the discovery sample set. Cross validation can reduce the bias but it usually does not remove the bias completely. Samples from ovarian cancer patients, especially early stage patients, are difficult to obtain due to its very low incidence. It is likely that the authors’ best approach to address this issue is to acknowledge it with a caveat since I doubt they have access to an independent validation sample set.
They should state that the results they present are in the context of samples obtained at diagnosis usually due to symptoms, they are likely upwardly biased due to choosing the maximum across a large number of candidate combinations (and state how many combinations were evaluated), but they are to be viewed as a stepping stone to further work in samples obtained prior to diagnosis usually obtained during the conduct of screening trials and from which stronger claims about earlier detection can be substantiated.

We thank the reviewer for their valuable critique of this manuscript that is resulting in a much stronger work. We apologize for addressing the previous comments in a way that did not fully ameliorate the concerns. We hope that our modifications, detailed below, have satisfactorily done so.

Regarding the sensitivity at 98% specificity, we agree with the reviewer that the results of model performance in the optimization process may have random variations. However, throughout the model optimization processes, there is no evidence and/or artifacts that would lead to upwardly biased results. After cross validation, the optimal marker combination was then validated using a new sample set of cases and controls, in Figure 3e (the curve labeled Test). Although these patient samples were acquired from the same facility/medical center, they were independent from those used in the cross validation. Regarding the reviewer’s concern about the use of serum samples obtained from symptomatic individuals, we modified the discussion in the manuscript to address this issue, as shown below:

In Page 8,
“…specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, prediction outcomes may differ in the clinical screening setting in which specimens are obtained in asymptomatic individuals before clinical diagnosis. Further studies, in the context of screening the general population, are warranted to evaluate the ability of the technology to identify pre-invasive and early-invasive disease.”

#2. I am not convinced that ANOVA (with a multiple comparisons adjustment) is the best test for assessing whether the sensitivity has increased when ß is changed. There would appear to be multiple dependencies between the estimates of sensitivity at different ß whereas ANOVA assumes independence of the values at different ß. However this is not a critical point in the manuscript.
My preference for the statement in the abstract is to be clear about the clinical challenge in early detection of ovarian cancer rather than to use the vague term “not [] sufficient diagnostic power”. A clear statement would help the field focus on the appropriate challenge, namely lack of sensitivity for early stage disease, rather than the vague challenge of insufficient power.
“Serum biomarkers are widely used as diagnostic indicators, but many are not sensitive enough for early stage disease when used for screening. Ovarian cancer, for example, is challenging to diagnose, in part because the few established biomarkers for the disease, while providing sufficient specificity when

followed by ultrasound, do not provide sufficient sensitivity for early stage disease."

We agree that early-stage detection is an important challenge in the area of ovarian cancer detection but not the only challenge. We believe that this type of technology, which enables detection or monitoring of ovarian cancer would be important for several purposes, including the detection of STIC (pre-cancerous) lesions, the detection of minimal residual disease after treatment or during recurrence, and very small volume 'advanced stage' disease which is equally as curable as early stage disease (e.g., patient with small tubal cancer and one lesion in the peritoneum or omentum), as well as early-stage detection. We note this in the modification to the introduction section of the manuscript, pasted below:

Page 1,
"Major opportunities for improving patient outcomes from ovarian cancer include increasing the sensitivity of early-stage detection while maintaining high specificity and the detection of minimum-residual/low-volume disease in treated patients. Accurate detection of known analytes does not always confer high sensitivity and specificity for disease, however."

Additionally, some investigators doubt that it is possible to detect early-stage ovarian cancers in many patients due to the often-rapid spread to the peritoneum after oncogenesis. The detection of STIC lesions, to prompt preventive treatment, or minimum residual disease after treatment, may be better propositions to improve patient outcomes. Therefore, we believe that it is not appropriate to introduce this work focused solely on early-stage detection as the only clinical use of this technology/platform in the abstract. We agree, though, as the reviewer pointed out in Comment #1, this work could potentially be viewed as a stepping-stone to further work for clinical screening trials and early-stage detection. Thus, we made a more general statement in the abstract. We agree that the statement could better fit the context outlined above. We have modified the abstract and pasted these changes below:

In Abstract:
"Serum biomarkers are widely used in medical decision making, but in many cases are not sensitive and/or specific enough to be diagnostic, especially in cases of low volume disease. In the quest to develop a screening test for ovarian cancer, for example, the few established biomarkers do not enable sufficient identification of patients with early-stage disease to impact mortality."

#3. The Page 1 conclusion would be clearer with the addition of the underline phrase: but the benefit of screening is still elusive due to lack of sensitivity for early stage disease.

Addressed above in comment #2.

#4. The authors have explained their estimate of 87% in the latest revision. However they need to be cautious and acknowledge an unknowable upward bias, since the estimate is the maximum across a large number of candidate combinations albeit cross-validated estimates.

Addressed in comment #1

#5. The authors have changed the language but only partly address this concern. In my judgment, due to the upward bias of the estimate due to the choice of the maximum across many candidates, there is insufficient evidence to assert that the performance exceeds current best clinical screening test (87% vs 84% sensitivity). I would accept an assertion that the performance is similar to the current best clinical screening test.
It is widely acknowledged that it is easier to "detect" symptomatic cases compared to controls, as the authors do in this study, than to detect asymptomatic cases amongst an apparently healthy population

(screening). Therefore this sentence should be modified as underlined: However, considering the fact that specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, performance is likely to be lower in the clinical screening setting in which specimens are obtained in asymptomatic individuals before clinical diagnosis.

This is relevant to Comment #1. We understand the reviewer's concern and thus we modified the discussion as follows:

In Page 8,
"… specimens obtained from symptomatic individuals at diagnosis were used for the development and assessment of the technology, prediction outcomes may differ in the clinical screening setting in which specimens are obtained in asymptomatic individuals before clinical diagnosis. Further studies, in the context of screening the general population, are warranted to evaluate the ability of the technology to identify pre-invasive and early-invasive disease."