Biosensors and Bioelectronics 279 (2025) 117402

Contents lists available at ScienceDirect

Biosensors and Bioelectronics

journal homepage: www.elsevier.com/locate/bios

Machine learning and statistical classification in CRISPR-Cas12a diagnostic assays

Nathan K. Khosla [a,1], Jake M. Lesinski [a,1], Marcus Haywood-Alexander [b], Andrew J. deMello [a,*], Daniel A. Richards [a,**]

[a] Institute for Chemical and Bioengineering, ETH Zurich, Vladimir-Prelog-Weg 1, 8093, Zürich, Switzerland
[b] Institute of Structural Engineering, ETH Zurich, Stefano-Franciscini-Platz 5, 8049, Zürich, Switzerland

ARTICLE INFO

Keywords:
CRISPR-Cas
CRISPR diagnostics
Machine learning
Statistical classification

ABSTRACT

CRISPR-based diagnostics have gained increasing attention as biosensing tools able to address limitations in contemporary molecular diagnostic tests. To maximize the performance of CRISPR-based assays, much effort has focused on optimizing the chemistry and biology of the biosensing reaction. However, less attention has been paid to improving the techniques used to analyze CRISPR-based diagnostic data. To date, diagnostic decisions typically involve various forms of slope-based classification. Such methods are superior to traditional methods based on assessing absolute signals, but still have limitations. Herein, we establish performance benchmarks (total accuracy, sensitivity, and specificity) using common slope-based methods. We compare the performance of these benchmark methods with three different quadratic empirical distribution function statistical tests, finding significant improvements in diagnostic speed and accuracy when applied to a clinical data set. Two of the three statistical techniques, the Kolmogorov-Smirnov and Anderson-Darling tests, report the lowest time-to-result and highest total test accuracy. Furthermore, we developed a long short-term memory recurrent neural network to classify CRISPR-biosensing data, achieving 100 % specificity on our model data set. Finally, we provide guidelines on choosing the classification method and classification method parameters that best suit a diagnostic assay's needs.

1. Introduction

Though most known for their role in genetic engineering, Clustered Regularly Interspaced Short Palindromic Repeats–CRISPR-Associated Protein (CRISPR–Cas) systems are increasingly employed as tools for biosensing and in vitro diagnostics (IVDs)(Kaminski et al., 2021; Suea-Ngam et al., 2020). Their utility in this regard can be attributed to their ability to specifically target exogenous nucleic acid sequences, such as those associated with invasive pathogens. Since their introduction, CRISPR–Cas-based biosensors have been applied to various diseases (Khosla et al., 2022), including cancers(Palaz), metabolic disorders(Ma et al., 2024), neurological conditions(Hajian), infectious diseases (Arizti-Sanz; Chen et al., 2018; Gootenberg et al., 2017), and cardiovascular disease(Chen et al., 2022). Success in each area was made possible through iterative optimization of assay chemistries and workflows(Chen et al., 2024; Del Giovane et al., 2024). Beyond assay optimization, Cas enzymes themselves have been engineered to enhance target specificity(Kaminski et al., 2021), catalytic activity(Gootenberg; Mahas et al., 2022), or thermal stability(Nguyen et al., 2023; Tian et al., 2020). Furthermore, novel chemical- and nanomaterial-based reporters have been synthesized to increase signals(Green et al., 2022; Yang et al., 2024), reduce background(Lesinski; Liu et al., 2022), or simplify assay readout(Broto; Xiong et al., 2021). Significantly less effort has been devoted to optimizing how the data emanating from these biosensors is analysed. This is surprising, given that key assay performance characteristics parameters, such as sensitivity, specificity, and time-to-result are known to be strongly influenced by data analysis methods (Fozouni; Lesinski et al., 2024a). This is particularly important when employing data processed in real-time, such as that commonly generated from CRISPR–Cas-based biosensors(Fozouni). Furthermore, although machine learning (ML) is well-established within the diagnostic community as a powerful tool for analyzing complex data sets

* Corresponding author.
** Corresponding author.
E-mail addresses: andrew.demello@chem.ethz.ch (A.J. deMello), daniel.richards@chem.ethz.ch (D.A. Richards).
[1] These authors contributed equally.

https://doi.org/10.1016/j.bios.2025.117402
Received 13 January 2025; Received in revised form 6 March 2025; Accepted 19 March 2025
Available online 29 March 2025

(Richens et al., 2020; Swanson et al., 2023), the combination of ML and CRISPR–Cas biosensing remains unexplored.

The rapid growth in CRISPR–Cas biosensing was catalyzed by two assays; the DNA Endonuclease-Targeted CRISPR Trans Reporter (DETECTR) assay and the Specific High Sensitivity Enzymatic Reporter UnLOCKing (SHERLOCK) assay(Abudayyeh; Chen et al., 2018). Whilst the intricacies of these assays differ, their mechanisms are conceptually similar. In both cases, a Cas protein is programmed to recognize, bind (according to 1:1 kinetics), and cleave a specific target DNA (*cis*-cleavage). After *cis*-cleavage, a conformational change occurs in the Cas protein, resulting in a loss of target specificity and the collateral cleavage of single-stranded DNA (*trans*-cleavage), a process which proceeds according to Michaelis-Menten kinetics (Fig. 1A). In the presence of a single-stranded DNA reporter containing a fluorophore and a quencher, both assays generate a fluorescence signal proportional to the *trans*--cleavage rate (and determined by the concentration of the initial target nucleic acid)(Chen et al., 2018; Gootenberg et al., 2018). Accordingly, both assays produce time-varying fluorescence signals that can be interpreted in real time. Similarly, to boost assay performance and decrease limits-of-detection, both assays are often coupled with Nucleic Acid Amplification Tests (NAATs).

Slope-based data analysis is the most common method for analyzing fluorescence data obtained from CRISPR–Cas biosensing assays (Fozouni; Pena et al., 2023; Ramachandran and Santiago, 2021). For example, Pena et al., developed a real-time method in which the first derivative of time *vs* fluorescence curves is analysed(Pena et al., 2023). In this method, a sample is deemed positive once the slope of the time varying fluorescence signal is greater than a pre-defined threshold (three standard deviations above the maximal negative slope) for three consecutive measurements. Additionally, Fozouni et al. used simple linear regression to determine the slope of fluorescence *vs* time curves, with a sample being deemed positive when its slope is greater than the average negative slope plus two standard deviations (95 % confidence interval - CI). Both approaches allow samples to be categorized in real time, while also decreasing the time needed for a positive diagnosis. This has important implications for point-of-care diagnostics, where time-to-result is often a critical performance metric(Atkinson et al., 2016).

Although effective in many situations, slope-driven methods are far from perfect. Numerical differentiations often simply describe a slope at a single point, and thus ignore the broader curve shape. Additionally, single-point methods are highly sensitive to noise, with minor signal fluctuations potentially leading to large variations in the local slope that can significantly impact the ability to precisely categorize samples(Pena et al., 2023). Linear regression methods that account for multiple data points circumvent this issue. However, linear regression often misdescribes non-linear data by forcing a linear fit upon the dependent and independent variables. Data from NAAT–CRISPR-Cas assays is the product of several non-linear processes (i.e. exponential amplification, 1:1 binding, Michaelis-Menten kinetics) (Fig. 1A). Thus, linearity is often only observed during a short portion of the assay. For high-titer samples, the linear phase is typically very short and thus normally described by a very small number of data points. Conversely, for low-titer samples, a linear fit is only possible after collecting many data points, requiring a longer time-to-result. As forcing a linear model onto non-linear data necessarily discards such non-linear effects, more sophisticated methods such as nonparametric statistics or ML are better suited for such analyses (Fig. 1B and C).

Although rarely used in CRISPR IVDs, nonparametric tests are routinely used to analyze and categorize real-time diagnostic data from lateral flow immunoassays (LFIAs)(Colombo et al., 2023). Nonparametric tests make no or minimal assumptions regarding the distribution

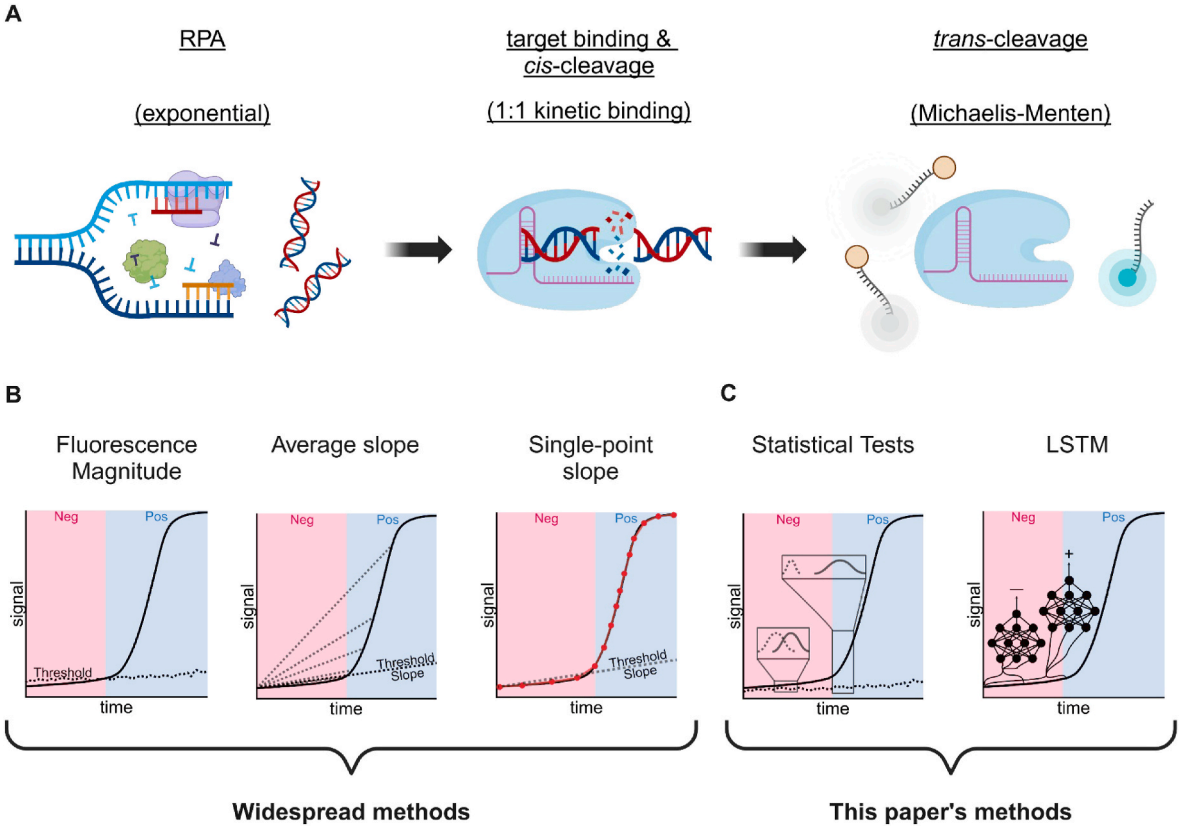


Fig. 1. NAAT–CRISPR–Cas biosensing and the classification algorithms evaluated in this work. A) Schematic describing the RPA–CRISPR–Cas12a reaction used to generate the data in this study. B) Graphical representations of Fluorescence magnitude, average slope, and single-point slope methods. C) Graphical representations of the non-parametric statistical tests and LSTM neural network presented in this work.

of the data. Thus, they can be more easily applied to smaller or less well-understood data sets. ML approaches have also gained traction in diagnostics, particularly those that employ recurrent neural networks (RNNs), a network architecture in which recent past events are included in computational decisions(McRae et al., 2022; Rahman et al., 2023). This capacity for "memory" enables classification and prediction from data trends rather than single data points. Such approaches have been effectively used in real-time NAATs, such as quantitative Polymerase Chain Reaction (qPCR) and quantitative loop-mediated isothermal amplification (qLAMP)(Sun et al., 2023; Waheed et al., 2022), and LFIAs (S. S. Lee et al., 2023; Turbé et al., 2021), but have yet to be integrated into CRISPR–Cas biosensors(Lee et al., 2022; Y. Y. Lee et al., 2023).

Herein, we use non-parametric statistical tests and a Long short-term Memory (LSTM) network to analyze time-varying fluorescence data from a CRISPR–Cas biosensing assay and show how these tools can be used to improve both the speed and accuracy of clinical sample classification. Specifically, we employ methods based on a family of non-parametric distribution goodness-of-fit statistics including the Kolmogorov-Smirnov (K-S) test, Anderson-Darling (A-D) test, and Cramér-von Mises (C-vM) test, as well as an LSTM network and apply these methods to data sets obtained from clinical HPV-16 vaginal swab samples. We show that non-parametric statistical methods and the LSTM network consistently outperform traditional analytical methods based on metrics such as sensitivity, specificity, and time-to-result (TTR). This work highlights the currently unexplored potential of integrating statistical analyses and ML into CRISPR–Cas biosensing systems.

2. Materials and methods

2.1. Fluorescence magnitude test

At each time point, analysis was performed without knowledge of any subsequent data. First, a time-dependent negative threshold was determined by taking the mean and standard deviation of the fluorescence data of all 24 negative trials at each time point. The cutoff was determined as the average of the negatives plus three standard deviations at a given time point. A sample was deemed positive when three consecutive fluorescence measurements exceeded the calculated threshold (a run length of 3).

2.2. Average slope test

At each time point, analysis was performed without knowledge of any subsequent data. First, a time-dependent negative average slope threshold was determined for each time, t_f, by performing a linear regression from t=0 to t=t_f on all the negative samples using the stats.linregress() function of the scipy.stats Python package. The threshold at each time was then calculated as the returned predicted slope plus three standard deviations of the predicted slope. For sample analysis, the same linear regression process was performed at each time point, and this value compared to the threshold value. A sample was deemed positive when three consecutive slope measurements exceeded the calculated thresholds, as described by Fouzoni et al.(Fozouni).

2.3. Single-point slope test

At each time point, analysis was performed without knowledge of any subsequent data. The maximum negative slope was determined by first finding the single point slope (change in fluorescence divided by change in time) for each measurement point in each of the 24 negative trials on the interval $t_0 = 0$ to $t_f = 90$ min and then taking the maximum observed value. The threshold was then calculated as this maximum plus three standard deviations of the set of all observed negative slopes. To analyze a sample, the single point slope (change in fluorescence dividing by change in time) was calculated at each time, and this value compared to the threshold value. A sample was deemed positive when three consecutive slope measurements (a run length of 3) exceeded the calculated thresholds, as described by Pena et al. (2023)

2.4. Kolmogorov-Smirnov test

At each time point, analysis was performed without knowledge of any subsequent data. This test compares two sample populations and returns a statistic indicating the likelihood that the two populations are from the same probability distribution. We constructed two populations; one from the sample to be analysed as well as one from the set of all negative sample readings. In both cases, the populations included only datapoints from the current time and two previous times (a sliding analysis window of length 3). In the case of the test to be analysed, this resulting population only contained data from that test, whereas the negative population contained datapoints within the sliding window for all 24 known negative tests. The two populations were then compared using a 2-sample, 2-sided Kolmogorov-Smirnov test ($\alpha = 0.003$) for goodness of fit from the ks_2samp() function of the scipy.stats Python package(Borovkov; Darling, 1957). Samples were identified as being positive after $p < \alpha$ once (i.e. requiring a run length of 1). This indicated that the sample was from a different distribution than the reference negative set, with associated probability of a false positive (type I error) equal to α. Cut-off values for run length, window length and α threshold were optimized to maximize total accuracy (Table S2, Figs. S2–4).

2.5. Anderson-Darling test

At each time point, analysis was performed without knowledge of any subsequent data. As above, two populations were constructed from the sample to be analysed as well as the set of all negative sample readings. The population to be analysed was determined for each time-point using a sliding window of length three. The same was then done for a combination of all negative tests. These two populations were then compared using a k-sample Anderson-Darling(Anderson and Darling, 1952) test for goodness of fit from the anderson_ksamp() function of the scipy.stats Python package with $\alpha = 0.0015$. The two populations were deemed to be from different distributions, indicating a positive sample, if $p < \alpha$ once (a run length of 1). Cut-off values for run length, window length and α threshold were optimized to maximize total accuracy (Table S2, Figs. S5–7).

2.6. Cramér-von mises test

At each time point, analysis was performed without knowledge of any subsequent data. As described above, two populations were constructed from the sample to be analysed as well as the set of all negative sample readings. The population to be analysed was determined for each time-point using a sliding window of length six. The same was then done for a combination of all negative tests. These two populations were then compared using a 2-sample, Cramér-von Mises(Cramér) test for goodness of fit from the cramervonmises_2samp() function of the scipy.stats Python package with $\alpha = 0.0001$. The two populations were deemed to be from different distributions, indicating a positive sample, if $p < \alpha$ once (a run length of 1). Cut-off values for run length, window length and α threshold were optimized to maximize total accuracy (Table S2, Figs. S8–10).

2.7. Long short-term memory network

An LSTM network is a specialised type of recurrent neural network designed to capture and retain information from sequential data. Here, the LSTM network was applied independently to the time-series fluorescence data from each trial. In training, the objective function in the form of binary cross-entropy loss, L, is minimised, for N sequences, i.e.

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)]$$

Where y_i is the true label of series i, $\hat{y}_i$ is the predicted probability of series i. To obtain representative values, data were averaged across all seeds. For each seed, data were split into training and validation sets, representing 70 % and 30 % of the full data, respectively. In this work, the LSTM RNN was built and trained using PyTorch(Paszke) with 2 layers of size 64, and a final sigmoid layer. The predicted label was set as positive (1) when the output of the final layer of the LSTM network exceeded the chosen threshold of 0.95 (Fig. S11). The models were trained with the Adam optimiser, with a learning rate of 5×10^{-5}. As the dataset is small, a dropout rate of 0.2 was chosen, along with an L2 regularisation on the parameters, as well as early stopping at 5000 epochs. Additionally, to analyze the effects of overfitting, 10 different seeds were used to split the dataset into training and validation sets, and 10 corresponding models trained and evaluated on each of these sets.

2.8. Acquisition of the model data set

For all analyses we used a previously published data set from Lesinski et al.(Lesinski). In this work the authors analysed 16 (8 positive and 8 negative, with triplicate assays of each sample) vaginal swabs for HPV-16. Specifically, we used the RPA-CRISPR Cas12a one-pot assay data. Briefly, the swabs were processed, concentrated, and analysed in an RPA-CRISPR Cas12a one-pot assay with fluorescence readout every 2 min. A in-depth description of the sample collection and preparation, as well as all assay procedures, can be found in the original work(Lesinski). This dataset included the RPA-CRISPR Cas12a raw fluorescence values and Anyplex qPCR values for the clinicals samples.

3. Results and discussion

3.1. Slope- and magnitude-driven classification

To establish appropriate benchmarks for binary classification in our CRISPR-Cas assay, we first evaluated the previously described slope-based methods reported by Pena et al. and Fouzoni et al. on the model data set(Fozouni; Pena et al., 2023). We also included a fluorescence magnitude analysis for comparison, as this is arguably the simplest method for classifying positives and negatives in analytical assays(Chen et al., 2018; Fozouni et al., 2021). We applied these widespread methods to previously published data obtained from an RPA–CRISPR-Cas assay for detecting HPV-16(Lesinski). The data set included time-varying fluorescence data from 24 positive and 24 negative assays, obtained using clinical vaginal swab samples, as confirmed by Allplex qPCR (Lesinski) (Table S1). From these analyses, we calculated key performance characteristics, namely sensitivity, specificity, and total accuracy (percentage of samples correctly indicated). These metrics are quantified in Fig. 2.

These results highlight several significant differences between the methods. The sensitivity of both the fluorescence magnitude and single-point slope approaches was poor, with respective true positive rates of 42 % and 38 %. Conversely, both tests displayed perfect specificity, yielding no false positives (100 % true negatives). The average slope method yielded a far greater sensitivity of 71 %, but was less specific, with a true negative rate of 88 %. The average slope algorithm demonstrated a better total accuracy. These data suggest that the fluorescence magnitude and single point slope algorithms are relatively strict, prioritising greater specificity at the expense of lower sensitivity. This could be important for situations in which false positives have a greater impact than false negatives, such as rare diseases with expensive and harmful treatments (e.g, rare cancers).

3.2. Non-parametric statistical classification

After establishing the performance of the fluorescence magnitude and slope-based analysis methods, we next evaluated the sensitivity, specificity, and total accuracy of nonparametric classification methods. We chose nonparametric methods since data from CRISPR–Cas assays cannot be assumed to originate from a known distribution. The K-S test determines the probability of two data sets being sampled from the same distribution by comparing their empirical cumulative distribution functions (ECDFs) (Fig. S1). The K-S test is a quadratic empirical distribution function (EDF) test, a family of statistical tests that also includes the C-vM test and the A-D test (Darling, 1957). Though conceptually similar, these tests diverge in how they compare the two distributions. The K-S test compares the maximum difference between the two distribution functions, making it less sensitive to deviations at the tails of the distributions. Conversely, the A-D test uses a weighted sum of the differences between the distribution functions, with more weight given to differences in the data at the extremes of the distributions. The C-vM test compares the squared differences between the distribution functions, integrated over the entire data range. Accordingly, it is sensitive to deviations across the entire distribution (Baumgartner and Kolassa). Given these differences, we decided to compare the performance of the three methods on our model data set (Fig. 3). In each case, empirical distribution functions were constructed from multiple consecutive measurements (window length). To improve robustness, samples were only classified after several consecutive distributions reached significance. We assessed multiple different

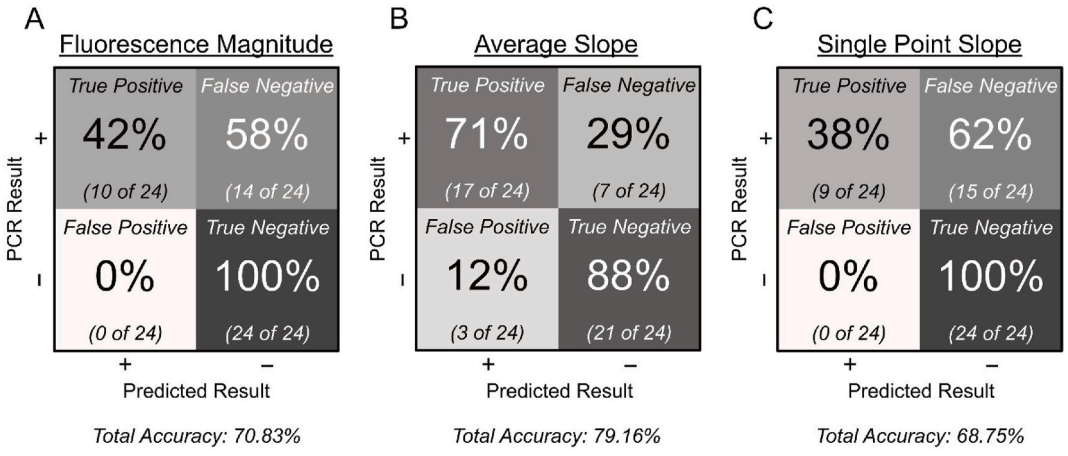


Fig. 2. Performance of widespread slope-based algorithms. A) Confusion matrix for the fluorescence magnitude method, analysed using a 99.7 % CI. B) Confusion matrix for the average slope method, analysed using a 99.7 % CI. C) Confusion matrix for the single point slope method, analysed using a 99.7 % CI.

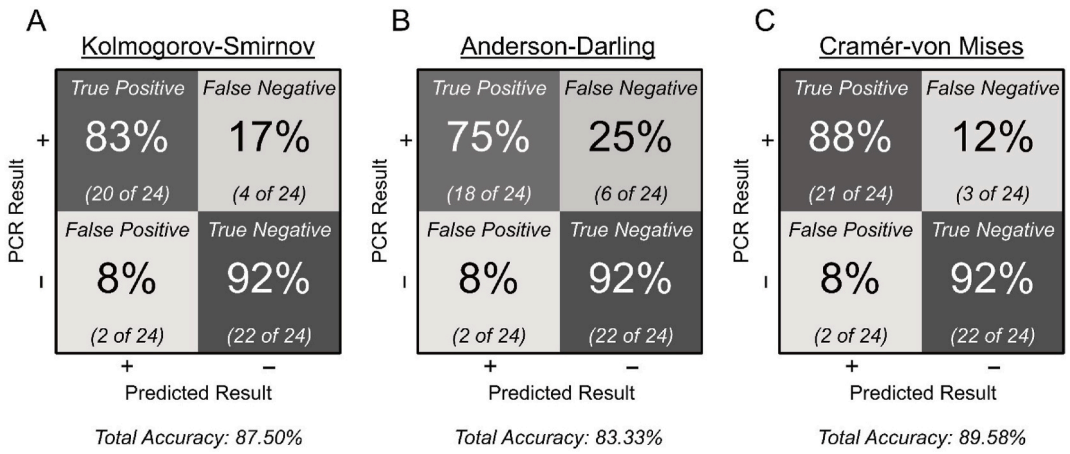
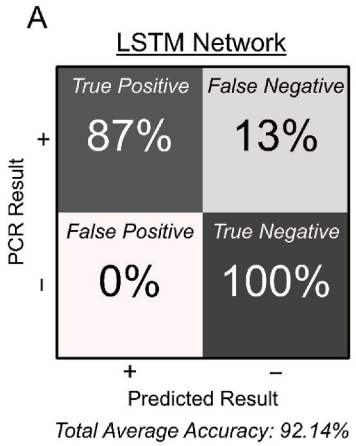


Fig. 3. Performance of quadratic EDF classification methods. A) Confusion matrix of the K-S test. Confidence interval = 99.7 %, window length = 3, run length = 1. B) Confusion matrix of the A-D test. Confidence interval = 99.85 %, window length = 3, run length = 1. C) Confusion matrix of the C-vM test. Confidence interval = 99.99 %, window length = 6, run length = 1.

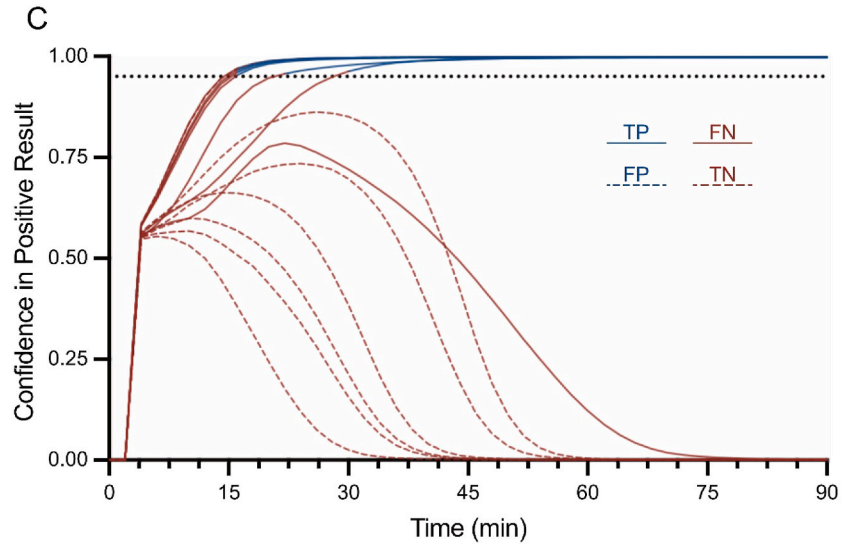
confidence intervals, window lengths, and run lengths (Figs. S2–S10), choosing the parameters that provide the greatest total accuracy.

As seen in Fig. 3, all three quadratic EDF classification methods were more sensitive than the slope-based methods, with true positive rates of 83, 75, and 88 % for K-S, A-D, and C-vM, respectively. However, this comes at the expense of specificity, as evidenced by the lower true



B

	True Positive	False Negative	False Positive	True Negative	Sensitivity	Specificity	Total Accuracy	F1 Score
Seed 1	6	2	0	6	75%	100%	86%	0.86
Seed 2	7	0	0	7	100%	100%	100%	1.00
Seed 3	6	3	0	5	67%	100%	79%	0.80
Seed 4	6	0	0	8	100%	100%	100%	1.00
Seed 5	6	0	0	8	100%	100%	100%	1.00
Seed 6	6	2	0	6	75%	100%	86%	0.86
Seed 7	7	1	0	6	88%	100%	93%	0.93
Seed 8	5	0	0	9	100%	100%	100%	1.00
Seed 9	7	1	0	6	88%	100%	93%	0.93
Seed 10	9	2	0	3	82%	100%	86%	0.90



D

	TTR Statistics (min)
	mean ± stdev
Seed 1	21.5 ± *9.8*
Seed 2	34.9 ± *20.8*
Seed 3	20.7 ± *15*
Seed 4	46 ± *15.7*
Seed 5	28.3 ± *17.8*
Seed 6	24 ± *10.8*
Seed 7	21 ± *2*
Seed 8	19.6 ± *4.5*
Seed 9	16.7 ± *7.8*
Seed 10	21.6 ± *11.5*

Fig. 4. Performance of an LSTM RNN. A) Confusion matrix. B) Sensitivity, specificity and accuracy of each LSTM network seed. C) LSTM network decision/confidence visualization over time for samples in a representative seed. The prediction threshold of 0.95 is indicated with a dotted black horizontal line. Solid lines indicate samples that are actual positives, while dashed lines indicate actual negatives. Samples classified as positive by the LSTM network are indicated in blue. This seed reported a single false negative, as shown by the solid red line. The LSTM network reported no false positives across all seeds. D) Average time-to-result of each LSTM network seed. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

negative rate of 92 % observed for all three methods. In terms of total accuracy, the quadratic EDF statistical tests significantly outperform both slope-based and fluorescence magnitude methods, with accuracies of 87.5, 83.33 and 89.58 %.

3.3. Classification using an LSTM RNN

With the growth in accessible computational power, ML methods have proven valuable in classification tasks(Sen et al., 2020; Singh et al., 2016). When considering time-varying data, recurrent neural networks are particularly relevant(Miao et al., 2015; Pawar et al., 2019). RNNs are a categorization tool where previous data information can be included in current decision-making processes. RNNs can handle such data due to their capacity for memory. Long-short-term-Memory RNNs employ three-way gating mechanisms (input, forget and output gates) to control the flow of information and capture long and short-term trends when making predictions. This is achieved using memory cells and hidden states, with one LSTM network cell being used as a node in the classic neural network configuration(HochreiterSchmidhuber). Data at each sequential point are passed into the LSTM network, along with the cell and hidden states of the previous sequential point. When an LSTM network is used for binary classification, the last layer comprises a sigmoidal activation function. Accordingly, the direct output of the LSTM network is a value between 0 and 1, representing the probability of the sequential data being the positive (1) class. A more detailed description of long short-term Memory RNNs is given by Staudemeyer and Morris(Staudemeyer). To test the utility of LSTM networks in classifying positive and negative samples in CRISPR–Cas assays, we established an LSTM network architecture, optimized the prediction threshold (Fig. S11), and applied it to our data set (Fig. 4a). Prior to training, the data set was randomly split into training and validation sets, the composition of which were dependent on the instantiated random number seed. This ensures a more accurate view of the performance of the LSTM network by observing the effect of different training set configurations. We observed different performance levels from different seeds and therefore different topologies of the training domain (Fig. S12). Accordingly, to avoid random bias, we used ten seeds to perform iterations of model training, validation, and data analysis (Table S3). Finally, we averaged the performance metrics of all validation sets from the ten models (Fig. 4a).

We applied the LSTM network to our data set and observed a sensitivity of 87 %, specificity of 100 % and total accuracy 92.14 % (Fig. 4b). Accordingly, the LSTM network outperforms both the slope-based and EDF statistical methods. However, the limited sample size used to train the network does increase the possibility of overfitting the LSTM network, distorting performance. This risk was mitigated by training in multiple independent trials with random seeds. An example of the decision process over time for one seed is shown (Fig. 4c). At each time point, the LSTM network analyses data from the start of the assay to that time point and returns a measure of confidence (between 0 and 1) that the data represents a positive test. Confidences above a threshold of 0.95 were accepted as positive tests, whereas confidence values less than 0.95 were classified as a negative test. Further, in assessing performance it is important to account for the variation introduced during the training process by comparing different models trained with different random seeds (Fig. 4D). In doing so, it becomes apparent that, although the assessment metrics may vary from each seed, the overall trend is toward improved performance, when comparing to classical statistical methods. However, it is important to note the significance of the data orientation on the results, highlighting the need for larger datasets when employing an LSTM network. Application of LSTM RNNs to other diagnostic systems would naturally require retraining to account for the specific features of each assay. That said, our data indicates that this effort will lead to improvements in test accuracy.

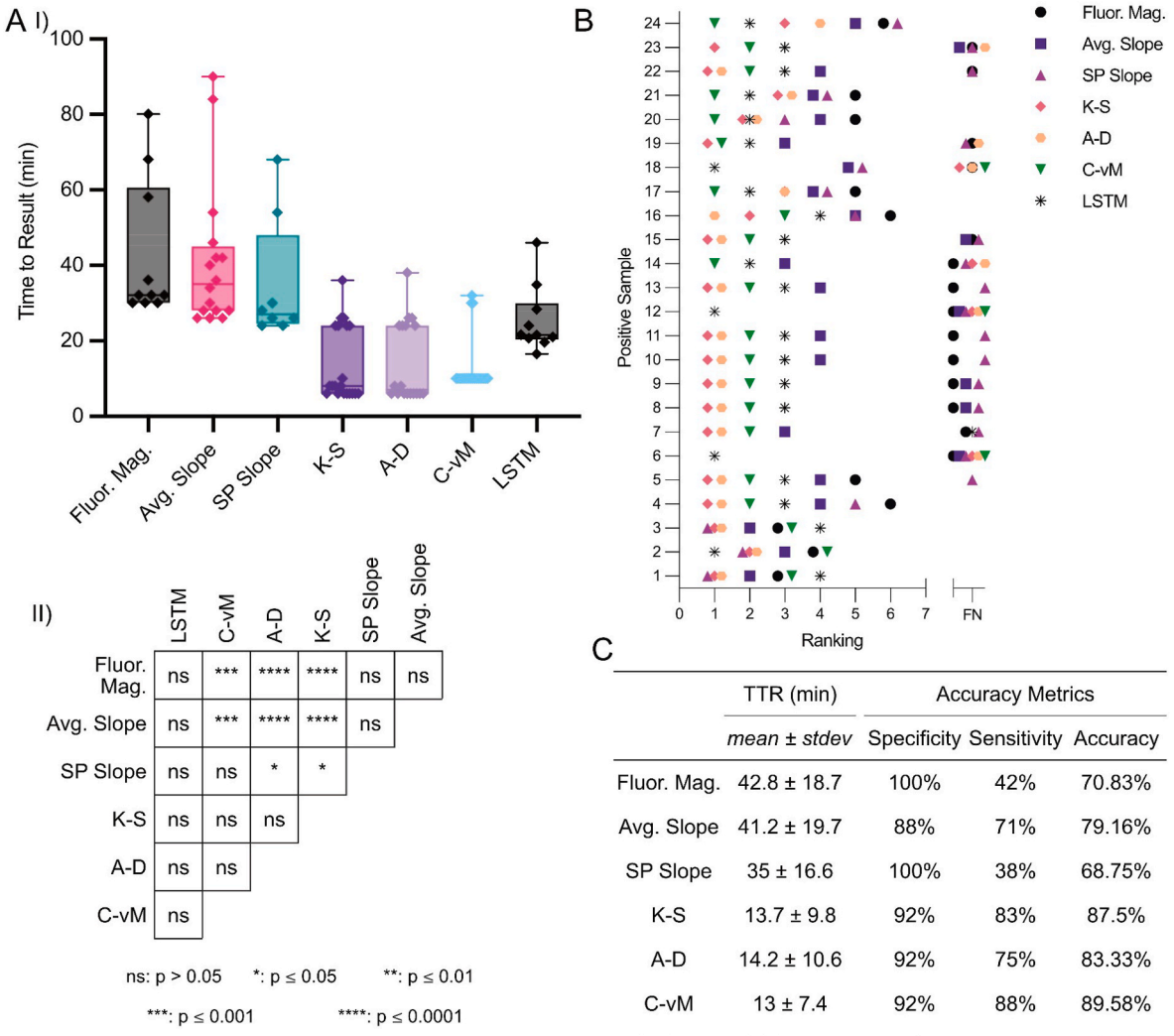
3.4. Impact of classification method on time-to-result

In addition to sensitivity, specificity, and accuracy, TTR is an important measure of assay performance. Obtaining a result within a specific time frame is paramount in many scenarios, such as rapid testing during an infectious disease epidemic/pandemic(Goldstein and Burstyn, 2020; Oeschger et al., 2021). Accordingly, we investigated how the different classification techniques impact TTR (Fig. 5A and B). Here, TTR is defined as the first time point at which a given sample matches the positive criterion, according to each individual method.

The data in Fig. 5 highlight several important features of each method. The slope-based methods returned the slowest average TTRs, ranging between 35.0 and 42.8 min. As shown in Fig. 5C, the range of TTRs for each method was large; 30–80 min, 26–90 min, and 24–68 min for the fluorescence magnitude, average slope, and single-point slope methods, respectively. The EDF methods returned the fastest average TTRs, (between 13 and 14.7 min), with ranges being 6–36 min, 6–38 min and 10–32 min for the K-S, A-D, and C-vM tests, respectively. The LSTM network returned an average TTR of 25.4 min. Notably, the LSTM network produced the smallest range, with TTRs between 16.5 and 46 min (note the fractional minute is an artifact of averaging a given sample's TTR from multiple seeds) (Fig. S11). Importantly, a statistically significant difference between these averages was only achieved between slope-based methods and the EDF methods, with similar methods (i.e. average slope *vs* single-point slope, or A-D *vs* K-S) generally displaying equivalent performance. We attribute this to the large variance in target titre between each patient sample. Considering the high sensitivities and average TTRs of 4.67, 4.67 and 6 min for K-S, A-D, and C-vM, respectively, it is clear that these methods prioritise rapid and accurate identification of positives. Despite the high confidence intervals (99.985 %, 99.97 %, and 99.99 % for the K-S, A-D, and C-vM respectively) utilized for the three methods, TTRs were significantly shorter than the benchmark methods. These performance metrics can be regulated by adjusting test parameters, such as confidence interval, window length and consecutive run length. Although a higher confidence interval makes the test "stricter" in that the bar for a sample being identified a positive is higher, this may increase TTR and lower sensitivity. Conversely, a shorter run length or lower confidence interval may perform well on sensitivity, catching even low-titer (and thus low-signal) tests quickly, but at the cost of specificity, where higher signal negatives are misidentified as positive. In this work, the high confidence interval was selected to maximize the total number of samples correctly identified in the data set. This highlights the importance of considering the clinical implications of false positives or false negatives for the particular diagnostic assay and setting the relevant parameters accordingly (Fig. S12).

Ranking the methods from lowest TTR (fastest) to highest TTR (slowest) for each sample (Fig. 5B) supports the trends observed for the average TTRs. The EDF methods are consistently fastest, followed by the LSTM network, then the slope-based methods, and finally the fluorescence magnitude method. In 20 out of 24 samples, the shortest (or joint-shortest) TTR was returned by an EDF test, with the Kolmogorov-Smirnov ranking 1st (or joint 1st) in 14 samples, the Anderson-Darling in 13, and the Cramèr-von Mises in ranking 1st in 6 samples. Interestingly, the four samples in which an EDF test did not provide the shortest TTRs are the ones in which no result was returned at all. In these cases, the LSTM network was the only method that correctly identified these positives, suggesting superior sensitivity. At the other extreme, the fluorescence magnitude method yielded the shortest TTR in 8 out of the 10 samples it correctly identified (out of a total of 24), suggesting this analysis method is both the slowest and least sensitive.

Fig. 5C provides a summary of the data obtained in this study and highlights the impact that different analysis methods have on performance metrics. By simply changing the analysis method, assay sensitivity ranged from 38 % to 100 %, specificity from 88 % to 100 %, and the average TTR from 4.7 min to 42.8 min. This strongly suggests that



	TTR (min)	Accuracy Metrics		
	mean ± stdev	Specificity	Sensitivity	Accuracy
Fluor. Mag.	42.8 ± 18.7	100%	42%	70.83%
Avg. Slope	41.2 ± 19.7	88%	71%	79.16%
SP Slope	35 ± 16.6	100%	38%	68.75%
K-S	13.7 ± 9.8	92%	83%	87.5%
A-D	14.2 ± 10.6	92%	75%	83.33%
C-vM	13 ± 7.4	92%	88%	89.58%
LSTM	25.4 ± 8.9	100%	87%	92.14%

Fig. 5. Comparison of TTR and accuracy metrics across all analysis methods. A) TTR values for tests analysed using the seven different methods (i) as well as pairwise comparisons for statistically significant differences in the TTR of each test(ii). B) Ranking of each method, from fastest (1) to slowest (7). False negatives are grouped as FN. C) TTR, specificity, sensitivity, and accuracy for each test.

different classification approaches should be considered for different applications, depending on priorities. The Kolmogorov-Smirnov, Anderson-Darling, and Cramér-von Mises tests may be most appropriate if a highly sensitive and rapid assay is more important than assay specificity. For example, in screening of mild but infectious diseases, a test may optimize for rapid and accessible feedback, favouring false positives with a non-harmful intervention and the ability to retest over a false negative and disease spread. Conversely, the LSTM RNN approach may be more appropriate if specificity is more important than sensitivity and assay time, and enough data is available to allow representative training of the network. The LSTM network may be more relevant when diagnosing diseases such as cancer, where treatments can require the administration of potentially harmful therapies, a test should take a more conservative classification approach, minimizing false positives at all costs. Finally, our results highlight the impact of test parameters such as confidence interval, window length, and consecutive run length (number of consecutive calls necessary to trigger positive indication) on sensitivity, specificity, and time-to-result. Accordingly, these parameters should be carefully optimized for each individual assay.

4. Conclusions

Whilst significant effort has focused on optimizing the chemistry and biology of CRISPR–Cas biosensing reactions, the analysis of CRISPR–Cas biosensing data has received relatively little attention despite its essential role within the IVD workflow. This work illustrates that statistical and ML methods have advantages in speed and accuracy over traditional fluorescence magnitude and slope-based approaches. Specifically, the K-S, A-D, and C-vM tests used in this work displayed optimal results across both performance metrics, having both the lowest TTR and the highest overall accuracy. The LSTM RNN, whilst being slightly less sensitive, displayed superior specificity than the statistical methods we tested, and outperformed traditional slope-based methods. Whilst we acknowledge the importance of increasing the relatively small data training datasets used in this work, we have shown that RNNs can be used as powerful classifiers in CRISPR–Cas diagnostics and that future application to large scale datasets has significant potential.

It is important to stress that we take no stance on the "best" optimization strategy. Rather, we simply aim to establish the idea that the choice of analysis method is equally as important as the inherent CRISPR–Cas biosensor performance. Considering the rapid growth of ML methods, we hope that this work sparks interest in using other ML

methods for the analysis and classification of CRISPR–Cas (and other) biosensing data. One potential avenue in this regard would be to explore different approaches for different aspects of assay classification (i.e. one method for classifying positive *vs* negative, and another for quantification of analyte concentration), and then employing these in parallel. Further avenues involve employing a mixture of expert architecture, a technique recently employed in the field of large language models to much success(Dong et al., 2024; Li et al., 2024). In the context of CRISPR–Cas biosensors, this could involve employing a mixture where some models ("experts") excel in rapidly calling high-signal positives and others in differentiating low-signal positives *vs* negatives. At each time point a router network would select the experts most suited to analyze the input data. Regardless of the route, the data presented herein suggests that applying concepts from machine learning to CRISPR–Cas diagnostics represents a promising and direct route for improving performance. Importantly, though we evaluated these methods on a model data set obtained from a CRISPR–Cas assay, they could easily be adapted and applied to any continuous data sets. It is our hope that others adopt and adapt these methods in their own work to broadly improve the performance of CRISPR-Cas-based diagnostic assays.

CRediT authorship contribution statement

Nathan K. Khosla: Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jake M. Lesinski:** Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Marcus Haywood-Alexander:** Methodology, Investigation, Formal analysis, Data curation. **Andrew J. deMello:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Daniel A. Richards:** Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to acknowledge that Fig. 1 was created with BioRender.com. DAR acknowledges funding from the ETH Career Seed Grant (SEED-13 21-2).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bios.2025.117402.

Data availability

Data will be made available on request.

References

Abudayyeh, O.O., Gootenberg, J.S., Essletzbichler, P., Han, S., Joung, J., Belanto, J.J., Verdine, V., Cox, D.B.T., Kellner, M.J., Regev, A., Lander, E.S., Voytas, D.F., Ting, A. Y., Zhang, F., 2017. Nature 550, 280–284. https://doi.org/10.1038/nature24049.

Anderson, T.W., Darling, D.A., 1952. Ann. Math. Stat 23, 193–212.

Arizti-Sanz, J., Freije, C.A., Stanton, A.C., Petros, B.A., Boehm, C.K., Siddiqui, S., Shaw, B.M., Adams, G., Kosoko-Thoroddsen, T.-S.F., Kemball, M.E., Uwanibe, J.N., Ajogbasile, F.V., Eromon, P.E., Gross, R., Wronka, L., Caviness, K., Hensley, L.E., Bergman, N.H., MacInnis, B.L., Happi, C.T., Lemieux, J.E., Sabeti, P.C., Myhrvold, C., 2020. Nat. Commun. 11, 5921. https://doi.org/10.1038/s41467-020-19097-x.

Atkinson, L.M., Vijeratnam, D., Mani, R., Patel, R., 2016. Int. J. STD AIDS 27, 650–655. https://doi.org/10.1177/0956462415591414.

Baumgartner, D., Kolassa, J., 2023. Commun. Stat. Simulat. Comput. 52, 3137–3145. https://doi.org/10.1080/03610918.2021.1928193.

Borovkov, A.A., 2004. Russ. Math. Surv. 59, 91–102. https://doi.org/10.1070/RM2004v059n01ABEH000702.

Broto, M., Kaminski, M.M., Adrianus, C., Kim, N., Greensmith, R., Dissanayake-Perera, S., Schubert, A.J., Tan, X., Kim, H., Dighe, A.S., Collins, J.J., Stevens, M.M., 2022. Nat. Nanotechnol. 17, 1120–1126. https://doi.org/10.1038/s41565-022-01179-0.

Chen, H., Li, Z., Chen, J., Yu, H., Zhou, W., Shen, F., Chen, Q., Wu, L., 2022. Bioelectrochemistry 146, 108167. https://doi.org/10.1016/j.bioelechem.2022.108167.

Chen, J.S., Ma, E., Harrington, L.B., Da Costa, M., Tian, X., Palefsky, J.M., Doudna, J.A., 2018. Science 360, 436–439. https://doi.org/10.1126/science.aar6245.

Chen, L., Hu, M., Zhou, X., 2024. Trends Biotechnol. 0. https://doi.org/10.1016/j.tibtech.2024.07.007.

Colombo, M., Bezinge, L., Tapia, A.R., Shih, C.-J., Mello, A.J., de, A., Richards, D., 2023. Sens. Diagn 2, 100–110. https://doi.org/10.1039/D2SD00197G.

Cramér, H., 1928. Scand. Actuar. J. 1928 13–74. https://doi.org/10.1080/03461238.1928.10416862.

Darling, D.A., 1957. Ann. Math. Stat. 28, 823–838. https://doi.org/10.1214/aoms/1177706788.

Del Giovane, S., Bagheri, N., Di Pede, A.C., Chamorro, A., Ranallo, S., Migliorelli, D., Burr, L., Paoletti, S., Altug, H., Porchetta, A., 2024. TrAC Trends Anal. Chem. 172, 117594. https://doi.org/10.1016/j.trac.2024.117594.

Dong, H., Chen, B., Chi, Y., 2024. https://doi.org/10.48550/arXiv.2404.01365.

Fozouni, P., Son, S., Díaz de León Derby, M., Knott, G.J., Gray, C.N., D'Ambrosio, M.V., Zhao, C., Switz, N.A., Kumar, G.R., Stephens, S.I., Boehm, D., Tsou, C.-L., Shu, J., Bhuiya, A., Armstrong, M., Harris, A.R., Chen, P.-Y., Osterloh, J.M., Meyer-Franke, A., Joehnk, B., Walcott, K., Sil, A., Langelier, C., Pollard, K.S., Crawford, E. D., Puschnik, A.S., Phelps, M., Kistler, A., DeRisi, J.L., Doudna, J.A., Fletcher, D.A., Ott, M., 2021. Cell 184, 323–333.e9. https://doi.org/10.1016/j.cell.2020.12.001.

Goldstein, N.D., Burstyn, I., 2020. Glob. Epidemiol. 2, 100031. https://doi.org/10.1016/j.gloepi.2020.100031.

Gootenberg, J.S., Abudayyeh, O.O., Kellner, M.J., Joung, J., Collins, J.J., Zhang, F., 2018. Science 360, 439–444. https://doi.org/10.1126/science.aaq0179.

Gootenberg, J.S., Abudayyeh, O.O., Lee, J.W., Essletzbichler, P., Dy, A.J., Joung, J., Verdine, V., Donghia, N., Daringer, N.M., Freije, C.A., Myhrvold, C., Bhattacharyya, R.P., Livny, J., Regev, A., Koonin, E.V., Hung, D.T., Sabeti, P.C., Collins, J.J., Zhang, F., 2017. Science 356, 438–442. https://doi.org/10.1126/science.aam9321.

Green, C.M., Spangler, J., Susumu, K., Stenger, D.A., Medintz, I.L., Díaz, S.A., 2022. ACS Nano 16, 20693–20704. https://doi.org/10.1021/acsnano.2c07749.

Hajian, R., Balderston, S., Tran, T., deBoer, T., Etienne, J., Sandhu, M., Wauford, N.A., Chung, J.-Y., Nokes, J., Athaiya, M., Paredes, J., Peytavi, R., Goldsmith, B., Murthy, N., Conboy, I.M., Aran, K., 2019. Nat. Biomed. Eng. 3, 427–437. https://doi.org/10.1038/s41551-019-0371-x.

Hochreiter, S., Schmidhuber, J., 1997. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Kaminski, M.M., Abudayyeh, O.O., Gootenberg, J.S., Zhang, F., Collins, J.J., 2021. Nat. Biomed. Eng. 5, 643–656. https://doi.org/10.1038/s41551-021-00760-7.

Khosla, N.K., Lesinski, J.M., Colombo, M., Bezinge, L., deMello, A.J., Richards, D.A., 2022. Lab Chip 22, 3340–3360.

Lee, S., Kim, S., Yoon, D.S., Park, J.S., Woo, H., Lee, D., Cho, S.-Y., Park, C., Yoo, Y.K., Lee, K.-B., Lee, J.H., 2023. Nat. Commun. 14, 2361. https://doi.org/10.1038/s41467-023-38104-5.

Lee, Y., Kim, Y.-S., Lee, D., Jeong, S., Kang, G.-H., Jang, Y.S., Kim, W., Choi, H.Y., Kim, J. G., Choi, S., 2022. Sci. Rep. 12, 1234. https://doi.org/10.1038/s41598-022-05069-2.

Lee, Y., Kim, Y.-S., Lee, D.I., Jeong, S., Kang, G.H., Jang, Y.S., Kim, W., Choi, H.Y., Kim, J. G., 2023. Viruses 15. https://doi.org/10.3390/v15020304.

Lesinski, J.M., Khosla, N.K., Paganini, C., Verberckmoes, B., Vermandere, H., deMello, A. J., Richards, D.A., 2024a. ACS Sens. 9, 3616–3624. https://doi.org/10.1021/acssensors.4c00652.

Lesinski, J.M., Moragues, T., Mathur, P., Shen, Y., Paganini, C., Bezinge, L., Verberckmoes, B., Van Eenooghe, B., Stavrakis, S., deMello, A.J., Richards, D.A., 2024b. Anal. Chem. 96, 10443–10450. https://doi.org/10.1021/acs.analchem.4c01777.

Li, J., Wang, X., Zhu, S., Kuo, C.-W., Xu, L., Chen, F., Jain, J., Shi, H., 2024. Wen, L. https://doi.org/10.48550/arXiv.2405.05949.

Liu, Q., Liu, M., Jin, Y., Li, B., 2022. Analyst 147, 2567–2574. https://doi.org/10.1039/D2AN00613H.

Ma, J., Zheng, Y., Xie, Y., Zhu, D., Wang, L., Su, S., 2024. Sens. Diagn 3, 1247–1252. https://doi.org/10.1039/D4SD00193A.

Mahas, A., Marsic, T., Lopez-Portillo Masson, M., Wang, Q., Aman, R., Zheng, C., Ali, Z., Alsanea, M., Al-Qahtani, A., Ghanem, B., Alhamlan, F., Mahfouz, M., 2022. Proc. Natl. Acad. Sci. USA 119, e2118260119. https://doi.org/10.1073/pnas.2118260119.

McRae, M.P., Rajsri, K.S., Alcorn, T.M., McDevitt, J.T., 2022. Sensors 22, 6355. https://doi.org/10.3390/s22176355.

Miao, Y., Gowayyed, M., Metze, F., 2015. 2015 IEEE workshop on automatic speech recognition and understanding (ASRU). Presented at the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 167–174. https://doi.org/10.1109/ASRU.2015.7404790.

Nguyen, L.T., Rananaware, S.R., Yang, L.G., Macaluso, N.C., Ocana-Ortiz, J.E., Meister, K.S., Pizzano, B.L.M., Sandoval, L.S.W., Hautamaki, R.C., Fang, Z.R.,

Joseph, S.M., Shoemaker, G.M., Carman, D.R., Chang, L., Rakestraw, N.R., Zachary, J.F., Guerra, S., Perez, A., Jain, P.K., 2023. Cell Rep. Med 4, 101037. https://doi.org/10.1016/j.xcrm.2023.101037.

Oeschger, T.M., McCloskey, D.S., Buchmann, R.M., Choubal, A.M., Boza, J.M., Mehta, S., Erickson, D., 2021. Acc. Chem. Res. 54, 3656–3666. https://doi.org/10.1021/acs.accounts.1c00383.

Palaz, F., Kalkan, A.K., Can, Ö., Demir, A.N., Tozluyurt, A., Özcan, A., Ozsoz, M., 2021. ACS Synth. Biol. 10, 1245–1267. https://doi.org/10.1021/acssynbio.1c00107.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S., 2019. In: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Pawar, K., Jalem, R.S., Tiwari, V., 2019. In: Rathore, V.S., Worring, M., Mishra, D.K., Joshi, A., Maheshwari, S. (Eds.), Emerging Trends in Expert Applications and Security. Springer, Singapore, pp. 493–503. https://doi.org/10.1007/978-981-13-2285-3_58.

Pena, J.M., Manning, B.J., Li, X., Fiore, E.S., Carlson, L., Shytle, K., Nguyen, P.P., Azmi, I., Larsen, A., Wilson, M.K., Singh, S., DeMeo, M.C., Ramesh, P., Boisvert, H., Blake, W.J., 2023. J. Mol. Diagn. 25, 428–437. https://doi.org/10.1016/j.jmoldx.2023.03.009.

Rahman, S.M.A., Ibtisum, S., Bazgir, E., Barai, T., 2023. Int. J. Comput. Appl. 185, 10–17. https://doi.org/10.5120/ijca2023923147.

Ramachandran, A., Santiago, J.G., 2021. Anal. Chem. 93, 7456–7464. https://doi.org/10.1021/acs.analchem.1c00525.

Richens, J.G., Lee, C.M., Johri, S., 2020. Nat. Commun. 11, 3923. https://doi.org/10.1038/s41467-020-17419-7.

Sen, P.C., Hajra, M., Ghosh, M., 2020. In: Mandal, J.K., Bhattacharya, D. (Eds.), Emerging Technology in Modelling and Graphics. Springer, Singapore, pp. 99–111. https://doi.org/10.1007/978-981-13-7403-6_11.

Singh, A., Thakur, N., Sharma, A., 2016. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). Presented at the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 1310–1315.

Staudemeyer, R.C., Morris, E.R., 2019. https://doi.org/10.48550/arXiv.1909.09586.

Suea-Ngam, A., Bezinge, L., Mateescu, B., Howes, P.D., deMello, A.J., Richards, D.A., 2020. ACS Sens. 5, 2701–2723. https://doi.org/10.1021/acssensors.0c01488.

Sun, H., Jiang, Q., Huang, Y., Mo, J., Xie, W., Dong, H., Jia, Y., 2023. Biomed. Signal Process Control 83, 104721. https://doi.org/10.1016/j.bspc.2023.104721.

Swanson, K., Wu, E., Zhang, A., Alizadeh, A.A., Zou, J., 2023. Cell 186, 1772–1791. https://doi.org/10.1016/j.cell.2023.01.035.

Tian, Y., Liu, R.-R., Xian, W.-D., Xiong, M., Xiao, M., Li, W.-J., 2020. Int. J. Biol. Macromol. 147, 376–384. https://doi.org/10.1016/j.ijbiomac.2020.01.079.

Turbé, V., Herbst, C., Mngomezulu, T., Meshkinfamfard, S., Dlamini, N., Mhlongo, T., Smit, T., Cherepanova, V., Shimada, K., Budd, J., Arsenov, N., Gray, S., Pillay, D., Herbst, K., Shahmanesh, M., McKendry, R.A., 2021. Nat. Med. 27, 1165–1170. https://doi.org/10.1038/s41591-021-01384-9.

Waheed, W., Saylan, S., Hassan, T., Kannout, H., Alsafar, H., Alazzam, A., 2022. Sci. Rep. 12, 4132. https://doi.org/10.1038/s41598-022-07954-2.

Xiong, E., Jiang, L., Tian, T., Hu, M., Yue, H., Huang, M., Lin, W., Jiang, Y., Zhu, D., Zhou, X., 2021. Angew. Chem. Int. Ed. 60, 5307–5315. https://doi.org/10.1002/anie.202014506.

Yang, C., Du, C., Yuan, F., Yu, P., Wang, B., Su, C., Zou, R., Wang, J., Yan, X., Sun, C., Li, H., 2024. Biosens. Bioelectron. 251, 116089. https://doi.org/10.1016/j.bios.2024.116089.