Different data analysis models for detecting miRNAs using nanopores

Caroline Koch[1‡], Seshagiri Sakthimani[2‡], Victoria Maria Noakes[2], Miruna Cretu[2], Richard Gutierrez[3], Mark Bruce[3], Julia Gorelik[4], Nadia Guerra[1], Joshua Edel[2], Aleksandar Ivanov[2]**

1 Department of Life Science, Sir Alexander Fleming Building, Imperial College London, South Kensington Campus, Imperial College Rd, London, SW7 2AZ, UK

2 Department of Chemistry, Molecular Science Research Hub, Imperial College London, White City Campus, 82 Wood Lane, London, W12 0BZ, UK

3 Oxford Nanopore Technologies, Oxford Science Park, Edmund Halley Road, Oxford OX4 4DQ, UK

4 National Heart and Lung Institute, ICTEM, Imperial College London, Hammersmith Campus, 72 Du Cane Rd, London, W12 0NN, UK

KEYWORDS Nanopores, Machine-learning, biomarker, miRNAs, data-analysis

ABSTRACT

Nanopore sensors generate complex ionic current traces that encode rich molecular information, making them well-suited for computational signal analysis. Building on previous work in which DNA-barcoded probes were developed for the multiplexed detection of biomarkers, this study explores different analytical approaches for interpreting binary nanopore signals. The method relies on identifying two classes of events: (1) non-delayed signals, indicating the absence of a target biomarker, and (2) delayed signals, resulting from probe-target interactions, proving the presence of the biomarker. To classify these events and enable robust biomarker detection, we applied three different computational strategies: (1) moving standard deviation (MSD), (2) spectral entropy (SE), and (3) a convolutional neural network (CNN) model. While MSD and SE represent conventional, rule-based approaches requiring manually defined thresholds, the CNN model offers a more scalable, automated solution for classifying nanopore events directly from raw time series data. All three methods were capable of distinguishing between signal classes; however, the CNN model demonstrated superior sensitivity and robustness. This work highlights the importance of data interpretation in nanopore sensing and presents a comparative framework for binary event classification. The findings pave the way for the development of machine learning-driven nanopore diagnostics capable of detecting diverse biomarker types at the single-molecule level.

INTRODUCTION

Nanopores are single-molecule electrical sensors capable of detecting and characterising individual molecules in real time[1]. The ability to distinguish and characterise molecules at single-molecule level is powerful, as individual molecules would normally be masked by ensemble averaging in conventional experimental techniques[2]. Due to this sensitivity, nanopores find use in various areas such as environmental monitoring[3], catalysis studies[4], DNA-based storage approaches[5,6], DNA and RNA sequencing experiments[7] and, most importantly for this study, the detection of biomarkers to differentiate between healthy and diseased patients[8,9].

The working principle involves applying an electrical bias across a nanoscale pore to generate a steady ionic current. When charged analytes translocate through the pore, the ion flow is temporarily disrupted, causing a characteristic drop in current. These current drops reveal information about the analyte's identity, structure, and physicochemical properties[1].

Due to the complexity of raw ionic current signals in nanopore experiments, a range of computational methods have been developed to extract the rich information encoded in the time series data. In nanopore sequencing, these current fluctuations enable real-time, base-by-base identification of nucleic acid sequences using machine learning algorithms[10]. In the context of biomarker detection, pattern recognition is critical for identifying characteristic translocation signatures. Two main challenges must be addressed: specificity and signal interpretation. Since nanopores lack intrinsic molecular specificity, it must be engineered, commonly through the use of DNA probes[11,12] or nanopore surface functionalisation[13] to selectively bind target biomarkers. Signal interpretation typically involves analysis of features such as dwell time and current amplitude[14], although more complex approaches also examine subpeak structures[15,16]. The optimal strategy depends heavily on the type of carrier used and the method by which specificity is introduced.

In our previous study, we developed a strategy for biomarker detection using DNA-barcoded probes and nanopore sequencing[17]. This method relies on sequencing a probe's barcode region and identifying a "delay" in the ionic current signal upon target binding, enabling single-molecule detection of biomarkers. Events were classified as either delayed or non-delayed, corresponding

to the presence or absence of a biomarker. By quantifying the fraction of delayed events, standard curves were generated for microRNAs (miRNAs), allowing for multiplexed detection of up to 40 miRNAs, as well as simultaneous detection of diverse analyte classes, including miRNAs, proteins, and small molecules. In this study, the moving standard deviation (MSD) was used to detect signal delays; however, this approach relied on manually defined thresholds, limiting sensitivity and scalability.

Building on this foundation, the present study evaluates three distinct data analysis strategies: (1) MSD, (2) spectral entropy (SE), and (3) a convolutional neural network (CNN) model to interpret nanopore signals. While MSD and SE represent traditional, rule-based thresholding methods, CNN embodies a modern, machine learning-based strategy that enables automated classification of time series data. This manuscript explores both paradigms: classical threshold-based classification and advanced deep learning, reflecting the shift toward data-driven methodologies in nanopore diagnostics. The performance of all methods was benchmarked against a conventional detection technique, namely RT-qPCR. This work introduces new analytical frameworks for interpreting binary nanopore signals and supports the development of automated, high-throughput diagnostic tools for single-molecule biomarker detection.

EXPERIMENTAL SECTION

Nanopore experiments

All nanopore experiments were performed with the MinION Mk1B device (ONT, UK) using MinION flow cells (R10.4.1, ONT, UK). The MinION Mk1B (ONT, UK) was connected through a USB 3.0 port to a computer with a minimum of 16 GB RAM and a graphics card. A membrane check was performed before each experiment using MinKNOW version 24.11.10 or above (ONT, UK).

DNA-barcoded probe and miRNA sequence

The DNA-barcoded probe was designed as previously described (Koch *et al.* Nature Nanotechnology, 2023). The DNA-barcoded probe sequence and its miRNA target (hsa-miRNA-27b-5p) were ordered from Integrated DNA Technologies (IDT) (Table 1).

Table 1: DNA-barcoded probe and hsa-miRNA-27b-5p sequence. 5Phos: Phosphorylation on 5'end, iSpC3: internal C3 spacer.

	Sequence (5'-3')
DNA-barcoded probe	5Phos/CCTAGTTCCGCTGGGATCGCTACGCCTTCGGCTCGTAATCATAGTCGAGT/iSpC3//iSpC3/GTTCACCAATCAGCTAAGCTCT
hsa-miRNA-27b-5p	AGAGCUUAGCUGAUUGGUGAAC

Sample preparation

The DNA-barcoded probe (0.225 µL, 20 µM) was incubated with 0.6 µL ligation c-strand (25 µM, 5' -CCC AGC GGA ACT AGG A -3') at room temperature (RT) for 1 h. Then, the mix was incubated with 7.5 µL ligation adapter (ONT, UK), 8.325 µL TA-ligase (M0367S, New England Biolabs, USA). The mix was centrifuged for 1 min at 3000 rpm and then incubated at RT for 20 min on a HulaMixer. Then, the DNA-barcoded probe was purified employing the SPRI (solid phase reversible immobilization) method using Ampure XP beads (A63880, Beckman Coulter, USA). Beads were added (23.3 µL), followed by 2x wash steps using 15 µL of short fragment buffer (ONT, UK). After the wash steps, the beads were resuspended in 20 µL elution buffer (ONT, UK), causing the DNA-barcoded probes to be released into the solution. Magnetic beads were then removed by a magnet, resulting in a clear solution of purified DNA-barcoded probes, ligated to the ligation adapter (ONT, UK).

Loading the flow cell

Before each run, the flow cell was flushed according to the manufacturer's protocol (ONT, UK). Then, the DNA-barcoded probe mix was prepared in a DNA lo-bind tube (Eppendorf, Germany) as follows: 37.5 µL of sequencing buffer (ONT, UK), 12 µL of ligated DNA-barcoded probes, and 25.5 µL of library solution (ONT, UK). For miRNA experiments, the miRNA was mixed with the DNA-barcoded probe at different concentrations ranging from 0 to 100 nM. The miRNAs were incubated for 30 minutes at RT before being added to the flow cell. All experiments were performed in triplicates.

Nanopore sequencing and basecalling

The software MinKNOW version 24.11.10 or above (ONT, UK) was used to record the data. For basecalling, the super high-accuracy algorithm was used (ONT, UK).

Barcode alignment

Each basecalled event was aligned to the reference barcode (GGGATCGCTACGCCTTCGGCTCGTAATCATAGTCGAGT). A local aligner with match score +5, mismatch score -4 and gap score -8 was used. Additionally, the events were further filtered using various parameters, namely i) minimum aligned threshold, ii) mismatch threshold, iii) initial sequence accuracy ("GGG"), and iv) mismatches in initial bases.

The minimum aligned threshold was reflected by the minimum number of bases that was matched with the reference (15 bases), whereas the mismatch threshold was defined as the maximum number of mismatches allowed (3 bases). The initial sequence accuracy parameter guaranteed alignment only at the start of the barcode sequence ("GGG"), rather than at random places within the sequence. The "mismatches in initial bases" extracts an initial number of bases and applies a mismatch threshold for them (3 mismatches in first 10 bases). The filtered events were then put through the classification algorithms for the delay to be identified.

Data analysis

Three methods for detecting the delay were tested, including 1) the MSD of the signal, 2) SE, and 3) a CNN model.

1. MSD

The first approach for detecting a delay in the electrical current signal was based on determining the MSD of the signal. The MSD is defined as:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left|x_{1-\mu}\right|^2}$$

where x is a vector containing N scalar observations and μ is the mean of x. Each standard deviation was calculated over a sliding window of length k across neighbouring elements of x_i. The MSD was calculated over k (number of bins), as well as other parameters outlined in Table 2. These parameters were used to classify an event as delayed or non-delayed.

Table 2: MSD thresholds. Standard MSD thresholds used in all experiments to determine whether an event was delayed or non-delayed.

	Parameter	Threshold
1	σ	< 0.003
2	Total number of bins	75
3	k (bins)	10
4	Delay fractional position	0.7

The parameters indicate 1) a minimum drop of the MSD to 0.003, 2) the total number of bins used for each event, 3) a length condition on the delay pattern, and 4) the fractional position of the delay within the MSD profile.

2. SE

The second approach for detecting a delay in the electrical current signal was based on determining the SE, which quantifies the power distribution within a signal by encoding the spectral attributes of the data. Using the power spectrum of a signal, SE is defined as:

$$\mathrm{SEN} = -\frac{\sum_{m=1}^{N} P(m)\log_2 P(m)}{\log_2(N)}$$

with

$$P(m) = \frac{S(m)}{\sum_{j=1}^{N} S(j)}, \sum_{m=1}^{N} S(m) = 1$$

where N is the number of frequency components, $S(m) = |X(m)|^2$ and X(m) is the discrete Fourier transformation of the signal x(n). The denominator, $\log_2(N)$, serves as a normalization factor,

ensuring that the SE is constrained within a range from zero to one. This normalization enables comparability between signals of varying lengths. It represents the maximum theoretical value of SE, achieved when all N frequencies contribute equally to the power spectrum. In the definition of SE provided, P(m) quantifies the percentage contribution of the mth frequency to the spectrum, essentially representing the probability that the signal contains that specific frequency. In the time domain, a flat signal corresponds to a power spectrum with few distinct frequency components, resulting in low SE values. Conversely, SE increases for signals with significant fluctuations in amplitude over time.

To compute SE in MATLAB, the pentropy function was used. This function operates on a vector x sampled at a predefined rate. It first calculates the spectrogram of the input time series and subsequently computes the Shannon entropy. Additionally, it returns instantaneous SEs, providing SE as a function of time. If a time-frequency spectrogram is known, denoted as S(t, f), the probability distribution of frequencies in the power spectrum can be described according to:

$$P(m) = \frac{\sum_t S(t,m)}{\sum_f \sum_t S(t,f)}$$

To compute instantaneous SE, the probability distribution at time t becomes:

$$P(t,m) = \frac{S(t,m)}{\sum_f S(t,f)}$$

Therefore, the normalised SE at time t reads as:

$$SEN_n(t) = -\frac{\sum_{m=1}^{N} P(t,m)\log_2 P(t,m)}{\log_2 N}$$

Instantaneous SEs were employed in this manuscript throughout. A SE threshold of 0.495 was used to differentiate delayed from non-delayed events, meaning that if the SE dropped below the value of 0.495, the event was classified as delayed.

3. Convolutional Neural Network

The data for training, testing and validation of the CNN model was selected from sequencing runs using various DNA-barcoded probe with varying miRNA concentrations (0nM, 1nM, 10nM, and

100nM) (Supplementary data 1). These reads were first aligned to its barcode sequence and then classified as delayed or non-delayed event based on the MSD method.

The classified events were plotted according to their class (delayed versus non-delayed). These plots were manually reviewed and selected. A total of 4500 images were curated, representing both classes (delayed versus non-delayed) and distributed across training, testing, and validation datasets as shown in Table 3.

Table 3: Distribution of delayed and non-delayed events for training, testing, and validation of the CNN model.

Class	Training	Testing	Validation
Delay	1000	1000	250
Non-Delay	1000	1000	250

The classifier architecture was inspired by the LeNet-5 convolutional neural network[18], originally designed for handwritten digit recognition. The model was implemented using PyTorch (v2.1.2). Different image resolutions were evaluated to determine the optimal input dimension. Both the test and blind datasets were used to assess model performance. All statistical metrics, including accuracy, precision, recall, F1-score, and MCC, were calculated using scikit-learn (v1.5.2).

Data to compare models

An additional 2000 images (1000 delayed, 1000 non-delayed) were selected without using the MSD to test the model independently of the labelling pipeline (Supplementary data 1). This dataset was used to evaluate the viability of MSD and the generalisation performance of the model and was labelled as "Blind data."

Statistical metrics

Recall measures the model's ability to correctly identify true positive (TP) events, which were events that were truly delayed. It is defined as:

$$\text{Recall} = \frac{\text{True positives (TP)}}{\text{True positives (TP)} + \text{False Negatives (FN)}}$$

where FN refers to false negatives, or delayed events that were incorrectly classified as non-delayed.

Precision measured the correctness of a model's prediction and is defined as:

$$\text{Precision} = \frac{\text{True positives (TP)}}{\text{True positives (TP)} + \text{False Positives (FP)}}$$

A high precision score indicated that the models produced few FPs and was conservative in assigning the "delay" label.

F1 Score is defined as the harmonic mean between precision and recall, which provides a single score that balances both sensitivity and correctness. It is calculated according to:

$$\text{F1 Score} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision} + \text{Recall}}$$

A high F1-score indicated that the model captured true delay events (TP) but also avoided excessively falsely detecting analytes (FP).

Mathews Correlation Coefficient (MCC) considers all elements of the confusion matrix (TP, TN, FP, and FN) and is defined as:

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FN})(\text{TN} + \text{FN})}}$$

MCC is considered one of the most reliable metrics for evaluating binary classification, especially when the classes are balanced.

Area under the ROC curve (AUC-ROC) assess the model's discriminatory power independent of a specific decision threshold. This metric evaluates the true positive rate (TPR) against the false positive rate (FPR) across all thresholds and is defined as:

$$\text{AUC} = \int_0^1 \text{TPR(FPR)d(FPR)}$$

An AUC of 1 indicates perfect separation between classes, while 0.5 indicates random guessing.

Accuracy is a common metric and is defined as the ratio of correctly classified instances to the total number of predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

RT-qPCR

Input RNA was prepared using 1-10 ng of total RNA per 15 µL RT reaction. The components of the RT kit (TaqMan™ MicroRNA Reverse Transcription Kit, 4366596, ThermoFisher, USA) were thawed and kept on ice. The RT primer (TaqMan™ Micro RNA Assay, 002174, ThermoFisher, USA) was thawed on ice, vortexed briefly, and then centrifuged to collect the contents at the bottom of the tube. The RT reaction mix was prepared according to the manufacturer's instructions. The RT reaction mix was then centrifuged and placed on ice. Subsequently, 7 µL of RT reaction mix was combined with 5 µL of total RNA in a reaction tube (MicroAmp™ Fast 8-Tube Strip, 0.1 mL, 4358293, ThermoFisher, USA). The content of the tube was mixed and centrifuged. Finally, 3 µL of 5xRT Primer (TaqMan™ Micro RNA Assay, 002174,, ThermoFisher, USA) were added to the reaction tube, centrifuged, and placed on ice before performing reverse transcription. To perform the RT reaction, the reaction tubes were placed into a PCR machine (StepOnePlus™, ThermoFisher, USA), then incubated using standard cycling, a reaction volume of 15 µL, and the following the manufacturers settings.

To prepare the PCR reaction mix, the PCR Master Mix (TaqMan™ Fast Advanced Master Mix, no UNG, A44359, ThermoFisher, USA) was thawed and mixed thoroughly. Then, the PCR Reaction mix was prepared according to manufacturer's instructions (1 µL TaqMan Small RNA Assay (20x), 10 µL PCR Master mix, and 4 µL nuclease-free water). 15 µL of PCR Reaction mix was transferred into each well of an optical reaction plate (MicroAmp® Fast 96-Well Reaction Plate, 4346907, Applied Biosystems, USA). Then, 5 µL of cDNA template, or NFW for control, were added to each well. The plate was sealed with optical adhesive film and centrifuged briefly to bring the PCR Reaction mix to the bottom of the wells. The manufacturer's instructions were used to run the RT-qPCR. Experiments were analysed using StepOnePlus software (ThermoFisher, USA).

RESULTS AND DISCUSSION

Developing technologies for early disease diagnostics can follow two main strategies: enhancing sensor sensitivity or enabling the simultaneous detection of multiple biomarkers to improve disease specificity. In this work, we employ nanopores which are single-molecule sensors with exceptional sensitivity, to detect low-abundance biomarkers that are critical for accurate and timely diagnosis. In parallel, we implement a multiplexed detection approach by combining nanopore sequencing with DNA-barcoded probes that selectively bind to target biomarkers [17]. Each DNA-barcoded probe specifically binds to its target biomarker, such as a protein or miRNA, and is subsequently sequenced using the MinION nanopore sequencing device (Oxford Nanopore Technologies, UK). Biomarker detection is determined by the presence or absence of a characteristic delay in the electrical signal, resulting in a quantitative readout of the biomarker, ultimately indicating the absence or presence of disease in patients (Figure 1A). The DNA-barcoded probe consists of three functional regions: (1) an adapter to control translocation speed, (2) a unique 35-base barcode sequence, and (3) a target-binding region, which can be either a complementary miRNA sequence or an aptamer for protein recognition (Figure 1B). Binding of the target biomarker alters the translocation dynamics, producing a measurable delay in the electrical current (Figure 1C, D). To analyse these delay patterns, we developed and evaluated three analytical methods: (1) moving standard deviation (MSD), (2) spectral entropy (SE), and (3) a convolutional neural network (CNN). MSD and SE represent conventional, rule-based thresholding methods that classify events as delayed when the signal falls below a predefined threshold. In contrast, the CNN employs a machine learning approach that enables automated classification of time-series data into two categories. This manuscript explores both paradigms, classical threshold-based classification and modern deep learning, reflecting the ongoing shift toward data-driven methodologies in nanopore diagnostics.

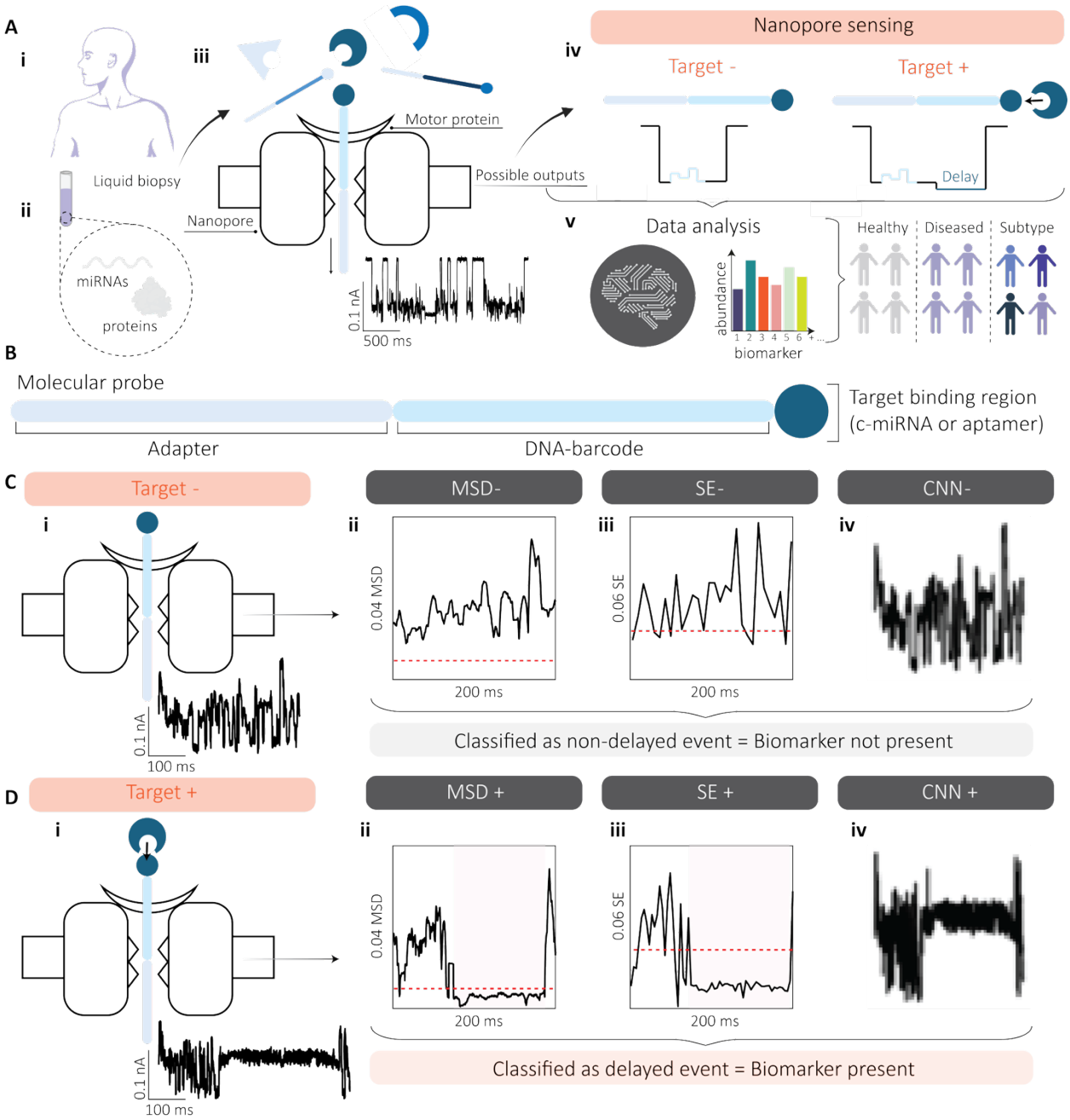

Figure 1 | Workflow for biomarker detection using a nanopore assay. (A, i) Liquid biopsies obtained from patients contain **(A, ii)** circulating biomarkers such as miRNAs and proteins. **(A, iii)** These biomarkers are hybridized with DNA-barcoded probes prior to nanopore sequencing. **(A, iv)** The resulting ionic current traces show no delay in the absence of the target biomarker but display a characteristic delay when it is present. **(A, v)** Computational analysis of the signal enables biomarker quantification and disease classification. **(B)** Schematic of the DNA-barcoded probe, comprising three regions: (1) the adapter, (2) a unique DNA-barcode sequence, and (3) the target-

binding region. **(C, i)** In the absence of the biomarker, the DNA-barcoded probe translocates through the nanopore without delay. The electrical signal is analyzed using three methods: **(C, ii)** moving standard deviation (MSD), **(C, iii)** spectral entropy (SE), and **(C, iv)** a convolutional neural network (CNN). Events that do not fall below the threshold or are classified as class 0 are identified as non-delayed. **(D, i)** In the presence of the biomarker, probe-target binding induces a delay during translocation. **(D, ii)** The MSD and **(D, iii)** SE methods detect a drop below the threshold, and **(D, iv)** the CNN classifies the event as class 1. These events are therefore considered delayed, indicating the presence of the biomarker.

To demonstrate the biomarker detection concept, previously developed DNA-barcoded probes[17] were employed to detect hsa-miR-27b-5p, serving as a proof-of-concept for evaluating different data analysis strategies. These DNA-barcoded probes generate two distinct signal types in nanopore measurements: (1) non-delayed events, indicating the absence of the target miRNA, and (2) delayed events, indicating its presence due to probe-target binding. As an initial analysis, the MSD of the ionic current time series was used to classify events into these two categories. The underlying hypothesis was that miRNA binding to the DNA-barcoded probe would prolong the occupancy of the nanopore by the probe-target complex, leading to reduced signal variability, observable as a decrease in the MSD of the ionic current. In all experiments, the DNA-barcoded probe was incubated with various concentrations of hsa-miR-27b-5p, as described in the Experimental section. The resulting data were used to optimize the MSD method for detecting current delays indicative of target miRNA presence. In brief, the MSD method divides the signal into discrete bins; if the current remains below a predefined threshold across a sufficient number of bins, the event is classified as delayed. The following analyses describe the calibration of these thresholds (Figure 2).

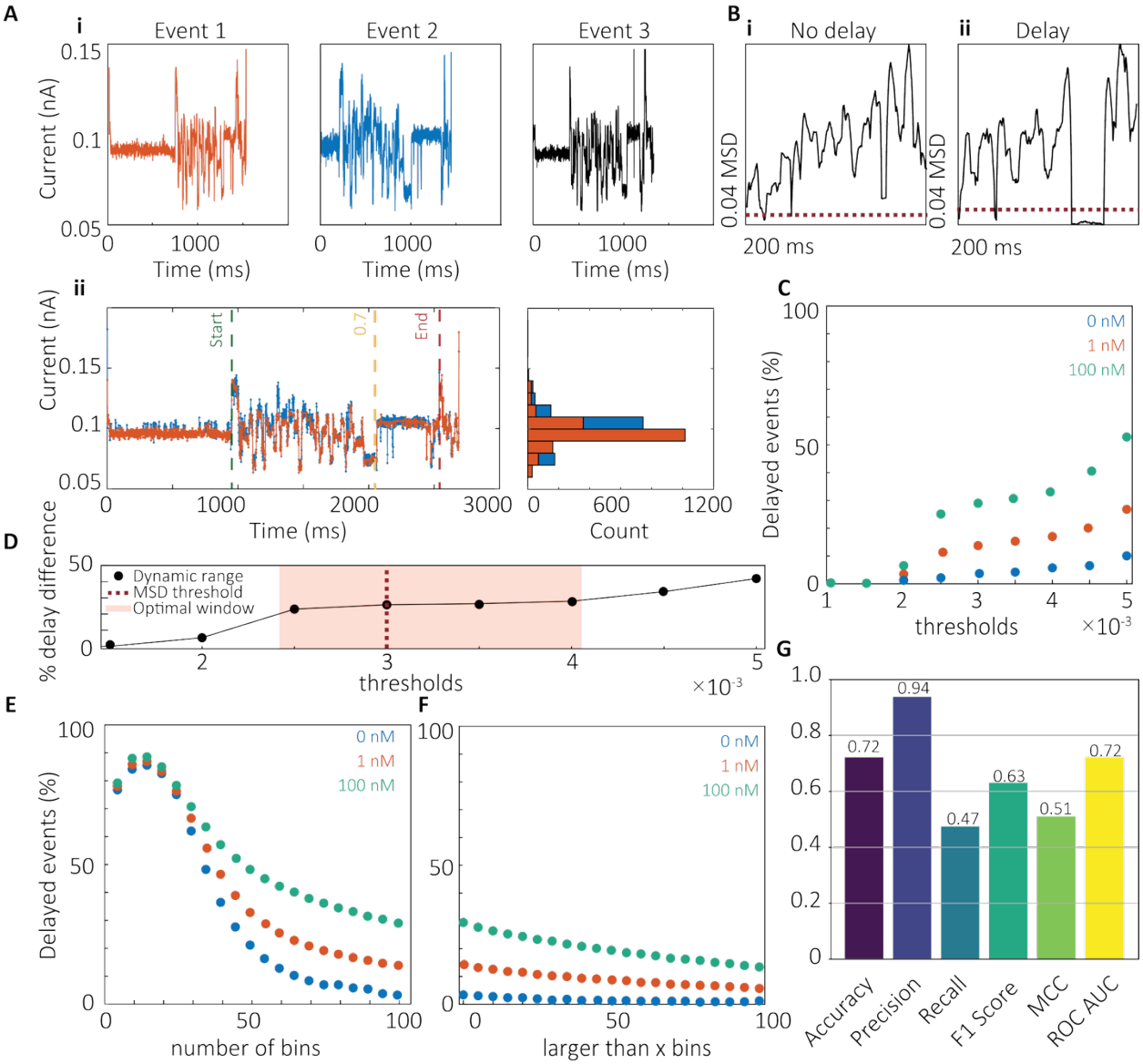

Figure 2 | Characterization of the MSD method for miRNA detection using DNA-barcoded probes. (A, i) Representative examples of three randomly selected delayed events, and **(A, ii)** an overlay of two events showing that the signal delay occurs at the fractional position 0.7. **(B, i)** Example of a non-delayed MSD signal and **(B, ii)** a delayed MSD signal, illustrating that delayed events exhibit a transient current drop below the defined threshold (0.003), whereas non-delayed signals remain above it. **(C)** Variation of the detection threshold from 0.001 to 0.005 at miRNA concentrations of 0, 1, and 100 nM. **(D)** Dynamic range of delay signal (delay at 100 nM - delay at 0 nM) indicating an optimal threshold between 0.0025 and 0.004. **(E)** Effect of increasing the number of bins from 5 to 100 while maintaining constant parameters (threshold = 0.003, fractional position = 0.7, minimum delay > 1 bin). **(F)** Effect of increasing the minimum delay from 0 to 100

bins while keeping other parameters constant (threshold = 0.003, fractional position = 0.7, 100 bins). **(G)** Performance metrics of the MSD method showing accuracy = 0.72, precision = 0.94, recall = 0.47, F1 score = 0.63, MCC = 0.51, and ROC AUC = 0.72.

The fractional position of the signal delay was first assessed by plotting three representative delayed events to visualise the delay pattern (Figure 2A, i). Dynamic time warping (DTW) was used to quantify signal similarity and locate the delay region (Figure 2A, ii). Delays consistently occurred in the final quarter of the signal, with DTW confirming strong similarity between events (Euclidean distance = 6.85), validating a fractional position threshold of 0.7 for delay detection. Representative MSD traces for non-delayed (Figure 2B, i) and delayed (Figure 2B, ii) events showed that only delayed events exhibited a transient drop below the threshold of 0.003. To determine the optimal threshold, three datasets (0, 1, and 100 nM hsa-miR-27b-5p) were analysed while varying the threshold from 0.001 to 0.005. Increasing the threshold led to a higher percentage of delayed events across all concentrations (Figure 2C). The dynamic range, which was defined as the delay at 100 nM minus the delay at 0 nM, was plotted to identify the range providing the greatest separation between control and miRNA-containing samples (Figure 2D). The plateau region, corresponding to thresholds between 0.0025 and 0.004, was selected as optimal. A threshold of 0.003 was chosen to maximise discrimination while maintaining a low false-positive rate in controls. Next, the influence of the *number of bins* and the *minimum delay length* ("larger than *x* bins") was examined. Varying the number of bins from 5 to 100 (while keeping other parameters constant: threshold = 0.003, fractional position = 0.7, minimum delay > 1 bin) revealed an inverse correlation between bin number and percentage delay (Figure 2E). A bin number of 75 was selected, as it produced consistently low delay percentages under control conditions. The *larger than x bins* parameter was then varied (threshold = 0.003, fractional position = 0.7, 100 bins) (Figure 2F). This parameter had little effect on the control data (0 nM) but markedly influenced delay detection in the presence of miRNA, as expected. To balance sensitivity and specificity, a threshold of *larger than 10 bins* was chosen, capturing most delayed events while tolerating minimal false positives. To validate the method, an independent dataset of 2,000 manually curated events (delayed and non-delayed) was analysed. The MSD classifier achieved an accuracy of 0.72, precision of 0.94, recall of 0.47, F1 score of 0.63, MCC of 0.51, and ROC AUC of 0.72 (Figure 2G). Further inspection revealed a subset of atypical events exhibiting high-

frequency noise within the delayed segment (Supplementary Data 2). These "noisy delays" elevated MSD values, leading to misclassification as non-delayed and contributing to a high false-negative rate. In diagnostic applications, such false negatives are particularly critical, as they may result in missed detections. To mitigate this, an alternative classification approach based on spectral entropy was developed to improve detection of delayed events and enhance overall accuracy of miRNA identification.

Spectral entropy (SE) was first introduced by Shannon in 1948 within the realm of communication theory, where it quickly expanded to fields such as pattern recognition[19]. It can detect potential anomalies or irregularities in time series data, by quantifying the level of disorder or unevenness in the frequency distribution and therefore was hypothesised to be able to differentiate between non-delayed and delayed events. To assess the feasibility of the SE approach, the same dataset previously analysed using the MSD method was employed. This dataset comprised events recorded from the DNA-barcoded probe after incubation with increasing concentrations of hsa-miR-27b-5p, as detailed in the Experimental section. The SE approach was optimised to ultimately analyse the electrical current traces for the presence or absence of a delay, to determine the presence of the miRNA target (Figure 3).

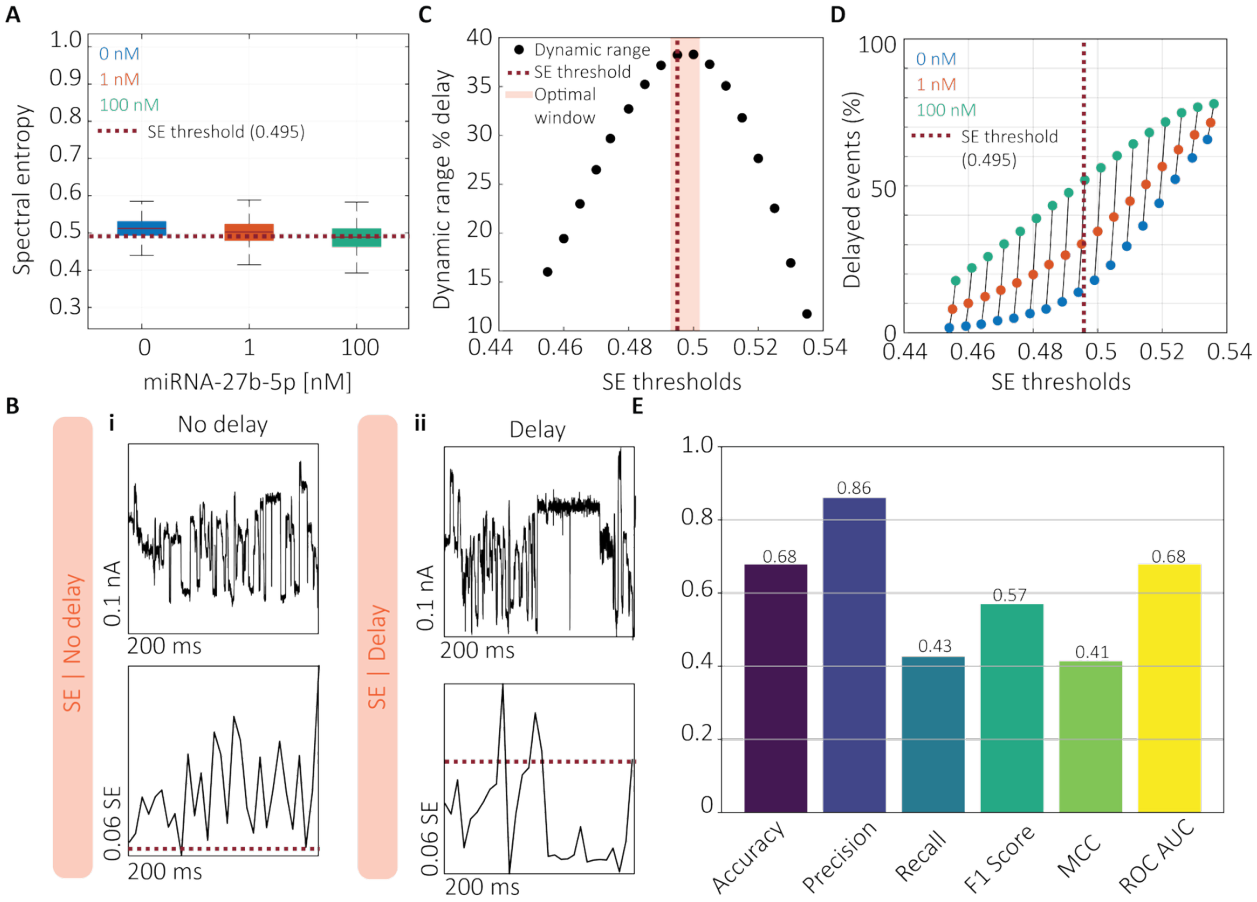

Figure 3 | Characterization of the SE method for miRNA detection using DNA-barcoded probes. (A) Distribution of SE values for events recorded without miRNA (control) and with increasing miRNA concentrations (1 nM and 100 nM). The classification threshold for delayed events was defined as the 25th percentile of the control distribution (SE = 0.495). **(B)** Example of a non-delayed event and **(B, ii)** a delayed event with their corresponding SE traces, showing that delayed signals drop below the SE threshold (0.495) whereas non-delayed signals remain above it. **(C)** Dynamic range of the delay signal (delay at 100 nM - delay at 0 nM) indicating an optimal threshold between 0.495 and 0.505. **(D)** Percentage of delayed events for data sets containing 0, 1, and 100 nM miRNA, calculated across varying SE thresholds. **(E)** Performance metrics of the SE method showing accuracy = 0.68, precision = 0.86, recall = 0.43, F1 score = 0.57, MCC = 0.41, and ROC AUC = 0.68.

To optimise the parameters used in the SE method, the first step was to determine an appropriate threshold for identifying delayed events (Figure 3A). SE distributions were plotted for three datasets (0, 1, and 100 nM hsa-miR-27b-5p). As expected, SE values decreased with increasing miRNA concentration. The 25th percentile of the control distribution (SE = 0.495) was selected as the classification threshold, separating delayed (SE < 0.495) from non-delayed (SE > 0.495) events. Representative electrical current traces and their corresponding SE profiles for a non-delayed (Figure 3B, i) and a delayed event (Figure 3B, ii) show that delayed signals exhibit a clear SE drop, consistent with the normalised SE formula and confirming SE as a robust indicator of delay. As previously established (Figure 2A, ii), miRNA-induced delays occurred within the region preceding the second C3 peak (fractional position > 0.7), and SE values from this region were therefore used for classification. The chosen threshold (SE = 0.495) was further validated by analysing the dynamic range of delay detection (delay at 100 nM - delay at 0 nM), which identified an optimal window between 0.495 and 0.505 (Figure 3C). Varying the SE threshold across the three datasets (0, 1, and 100 nM hsa-miR-27b-5p) revealed that the dynamic range increased up to the set threshold and declined thereafter (Figure 3D). This emphasises the importance of precise threshold selection, as inappropriate values may capture barcode-related variations rather than true delay signals. To evaluate the classifier's performance, the same dataset of 2,000 events previously analysed with the MSD method was re-evaluated using the SE approach (Figure 3E). The SE classifier achieved an accuracy of 0.68, precision of 0.86, recall of 0.43, F1 score of 0.57, MCC of 0.41, and ROC AUC of 0.68. Despite its utility, the SE method has several limitations. It is sensitive to spectral resolution effects, as the number of frequency components (N) used in SE computation strongly influences SE magnitude, complicating cross-dataset comparisons. This effect is particularly pronounced when signal patterns repeat or when sampling resolution is high, leading to length-dependent biases, where shorter signals tend to be classified as non-delayed, while longer ones appear delayed (Supplementary Data 3). A further limitation shared by both the MSD and SE methods is their dependence on manually defined thresholds. Threshold choice substantially affects the estimated proportion of delayed events (Figure 3D), reducing reproducibility and introducing user bias. To overcome this limitation, we developed an alternative, threshold-free classification strategy based on a convolutional neural network (CNN) inspired by the LeNet-5 architecture (Figure 4).

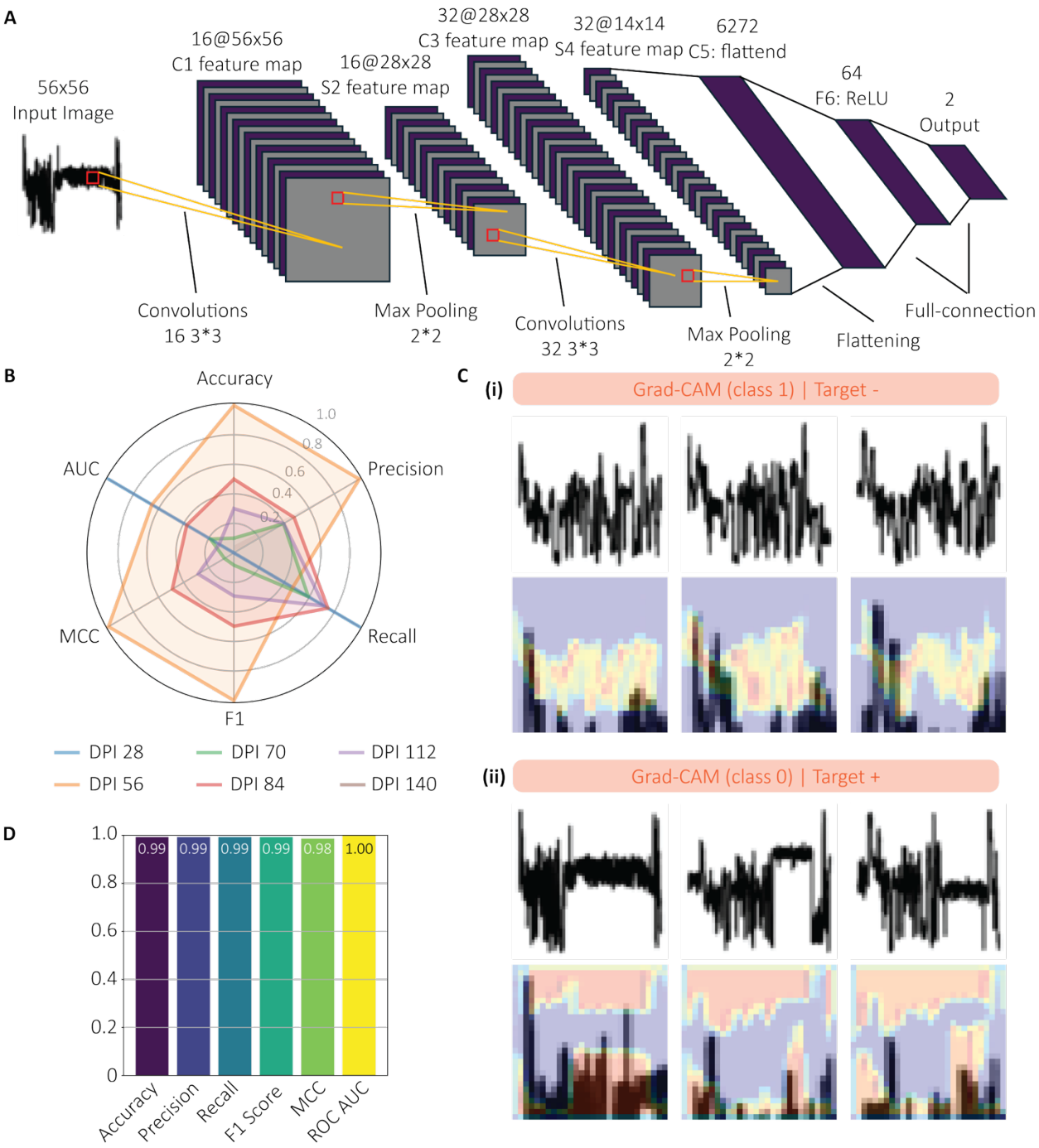

Figure 4 | Characterization of the CNN model for miRNA detection using DNA-barcoded probes. (A) Architecture of the CNN model inspired by the LeNet-5 design. **(B)** Radar plots showing normalized performance metrics (min-max scaling per metric) for CNN models trained at six input image resolutions (28, 56, 70, 84, 112, and 140 pixels), highlighting relative performance differences. **(C, i)** Representative non-delayed events and **(C, ii)** delayed events with

their respective Grad-CAM visualizations, illustrating feature activation within the delay region. **(D)** Performance metrics of the CNN model showing accuracy = 0.99, precision = 0.99, recall = 0.99, F1 score = 0.99, MCC = 0.98, and ROC AUC = 1.00.

The CNN model was designed to capture the unique characteristics of signals generated by DNA-barcoded probe translocations, both with and without miRNA binding. The final architecture (Figure 4A) comprised two convolutional layers with 16 and 32 filters, respectively, each followed by ReLU activation and 2 × 2 max pooling for spatial down-sampling. These layers enabled the extraction of local signal features such as amplitude shifts and contour shapes. The resulting feature maps were flattened and passed through a fully connected layer with 64 units before reaching the final classification layer, which outputs one of two classes: delay (class 0) or non-delay (class 1). To optimize input image resolution, CNN models were trained at six resolutions (28, 56, 70, 84, 112, and 140 pixels) and evaluated using normalized performance metrics (min-max scaling per metric; Figure 4B). The best performance was achieved at 56 DPI, which was therefore selected for subsequent analyses. This architecture was chosen for its simplicity, fast training time, and strong performance on low-resolution input images (Supplementary Data 4). Gradient-weighted Class Activation Mapping (Grad-CAM) was used to generate saliency maps to identify which regions of the input contributed most to the classification outcome. Grad-CAM was selected over Grad-CAM++ because it produced a smoother, more coherent envelope of the signal, especially for the non-delay representations (Supplementary Data 6). Original current traces were first corrected for signal deviation (Supplementary Data 5) and converted into 56 × 56 pixel grayscale images (Figure 4C, i-ii). These images were normalized using the mean and standard deviation from model training and then subjected to Grad-CAM analysis to generate heatmaps highlighting areas of model focus (black = no focus, blue = mean focus, red = highest focus; Figure 4C, i-ii and Supplementary Data 7). For non-delayed events, the model concentrated on sharp transitions and local fluctuations, whereas for delayed events, attention was focused on extended flat regions in the center of the trace, consistent with unzipping-induced delays. These patterns indicate that the CNN learned to distinguish delayed from non-delayed signals based on relevant biophysical features rather than extraneous differences such as barcode sequence. Grad-CAM overlays confirmed that the model's decisions were grounded in meaningful signal regions, supporting its interpretability and generalizability. To benchmark performance, the same dataset

of 2000 events previously analyzed with the MSD and SE methods was re-evaluated using the CNN (Figure 4D). The model achieved an accuracy of 0.99, precision = 0.99, recall = 0.99, F1 score = 0.99, MCC = 0.98, and ROC AUC = 1.00.

As a final step, the three different analysis methods (MSD, SE, and CNN) were compared based on their performance and were evaluated against RT-qPCR (Figure 5).

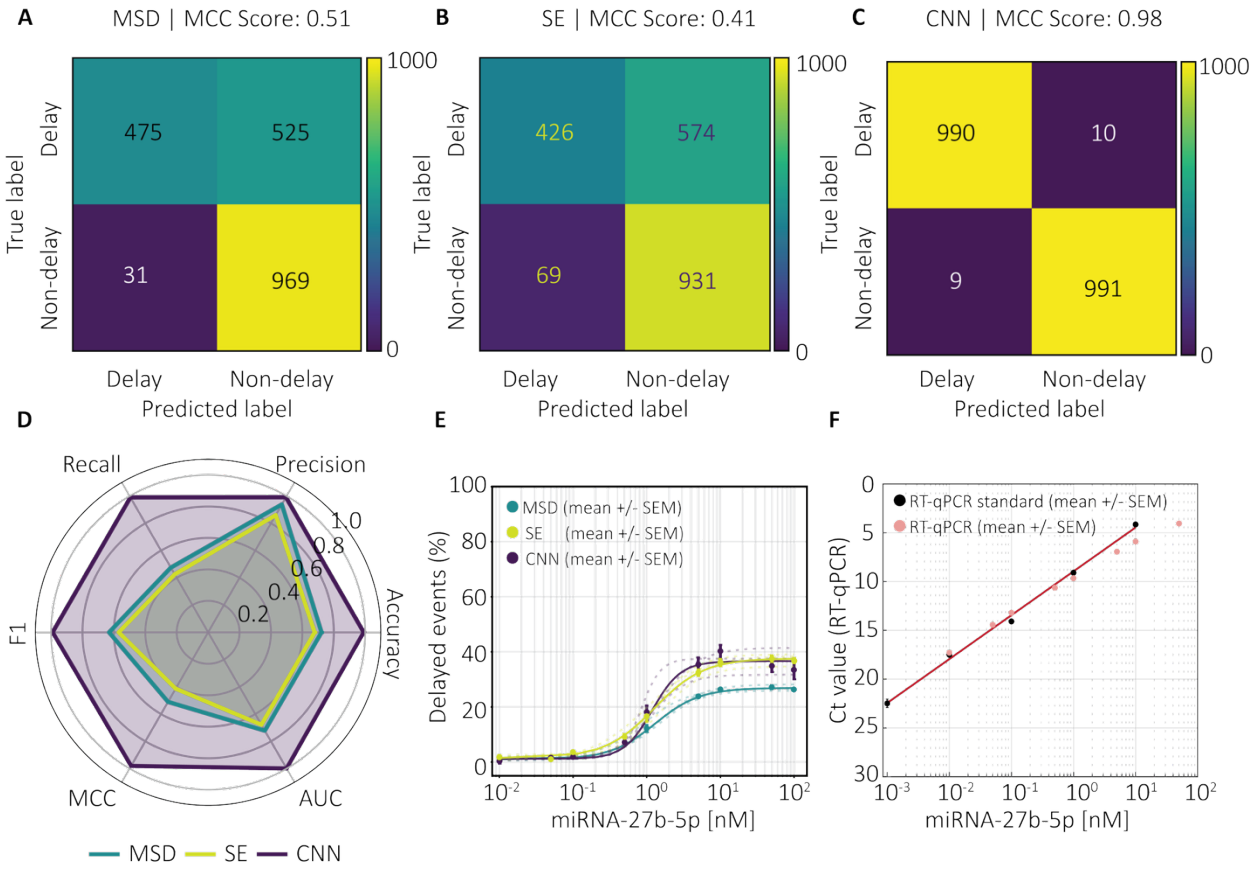

Figure 5 | Comparison of the MSD method, SE method, and CNN model for miRNA detection using DNA-barcoded probes. (A-C) Confusion matrices showing classification outcomes for delay prediction using the **(A)** MSD method, **(B)** SE method, and **(C)** CNN model. **(D)** Radar plots summarizing performance metrics (accuracy, precision, recall, F1 score, MCC, and ROC AUC) for all three approaches. **(E)** Concentration-percentage delay relationships for the DNA-barcoded probe in the presence of hsa-miR-27b-5p (0-100 nM), fitted using the Hill equation. The MSD method yielded n_h = 1.43, K_e = 1.30 nM, V_{max} = 26.88%, R^2 = 0.996; the SE method yielded n_h = 1.29, K_e = 1.25 nM, V_{max} = 37.33%, R^2 = 0.998; and the CNN model yielded n_h

= 2.25, K_e = 1.19 nM, V_{max} = 36.59%, R^2 = 0.965 (n = 3, $n_{total-events}$ = 1,642,070). Data are presented as mean ± SEM. All experiments were performed in sequencing buffer. **(F)** RT–qPCR analysis showing the standard curve and corresponding miRNA concentrations used for nanopore measurements, demonstrating decreasing cycle threshold (Ct) values with increasing miRNA concentration.

The confusion matrix revealed that the MSD method tends to miss true delays, reflecting low sensitivity but high specificity and precision (Figure 5A). Similarly, the SE method performs well in identifying non-delays yet frequently fails to detect true delays, as indicated by its moderate MCC value of 0.41 (Figure 5B). These findings suggest that both MSD and SE are prone to false negatives, underscoring a key limitation in sensitivity. In contrast, the CNN model exhibited a markedly improved balance between precision and recall, achieving near-perfect classification accuracy (Figure 5C). This strong performance indicates that the deep learning approach effectively captures complex signal patterns directly from raw traces, enabling more consistent and scalable delay detection than threshold-based methods. To quantitatively compare model performance, standard metrics such as accuracy, precision, recall, F1 score, MCC, and ROC AUC were evaluated (Figure 5D). The CNN model achieved values close to 1.0 across all parameters, demonstrating both exceptional predictive power and balanced classification. By contrast, the MSD and SE methods displayed lower and more variable performance, particularly in recall, confirming their tendency to under detect true delays. Collectively, these results highlight the superior performance of the CNN model for this classification task. Comparison of the concentration-percentage delay relationships further emphasized the differences among methods (Figure 5E). All three approaches yielded sigmoidal response curves, with percentage delay increasing in a concentration-dependent manner, consistent with expected binding kinetics. The CNN and SE methods showed enhanced sensitivity at lower miRNA concentrations and earlier saturation compared to MSD, suggesting improved capacity to detect subtle delay signals. The close overlap of fitted curves and low standard errors across replicates indicate high reproducibility and robustness of the nanopore-based detection strategy. Finally, RT-qPCR validation confirmed the concentration-dependent detection of hsa-miR-27b-5p (Figure 5F). A standard curve generated from a series of target concentrations (0-50 nM) showed the expected decrease in cycle threshold (Ct) values with increasing miRNA levels. The strong correspondence between nanopore-derived

delay metrics and RT–qPCR results support the accuracy and quantitative reliability of the nanopore assay for label-free miRNA detection.

CONCLUSION

This study presents a comparison of three approaches for classifying delay and non-delay events in nanopore signals generated using DNA-barcoded probes. In nanopore-based detection, the presence of a delay in the ionic current indicates target binding. Delayed events were identified using the MSD and SE method, which both relied on manual thresholding. While effective, both requires fine-tuning for each new analyte, limiting its scalability and introducing user bias.

To overcome these limitations, a lightweight CNN model, inspired by LeNet-5, was developed for robust, two-class classification of nanopore events. The model was trained on 5000 image representations of raw current traces, with 2000 manually curated for blind testing. Pre-processing steps removed extreme outliers and preserved signal dynamics within a 200–700 pA range. Various resolutions were evaluated, with the 56×56 format achieving the best performance across all metrics, including precision, recall, F1-score, AUC-ROC, and MCC. Grad-CAM visualisation confirmed that the CNN focused on plateau regions associated with delays, avoiding barcode-related signal regions. Comparative analysis showed that the CNN consistently outperformed MSD, particularly in sensitivity, validating its improved detection capability. However, the CNN model's generalisation is currently limited to patterns learned from specific probe configurations, such as delay events preceding the C3-peak. Future improvements could involve hybrid architectures combining CNNs with time-series models or transformer-based encoders, and real-time implementation for live signal classification.

In conclusion, this work demonstrates that nanopore-based detection combined with advanced computational analysis enables robust and quantitative miRNA profiling. Among the three evaluated models, the CNN approach achieved the highest accuracy and sensitivity, outperforming MSD and SE in classifying true delay events and capturing concentration-dependent signal changes. The strong agreement between nanopore-derived delay metrics and RT-qPCR standard curves further validates the method's quantitative reliability, confirming that nanopore signal delays faithfully reflect miRNA abundance across a wide dynamic range.

REFERENCES

1. Xue, L. *et al.* Solid-state nanopore sensors. *Nat Rev Mater* **5**, 931–951 (2020).

2. Miles, B. N. *et al.* Single molecule sensing with solid-state nanopores: novel materials, methods, and applications. *Chem Soc Rev* **42**, 15–28 (2012).

3. Lim, F. S. *et al.* Advancing pathogen surveillance by nanopore sequencing and genotype characterization of Acheta domesticus densovirus in mass-reared house crickets. *Scientific Reports 2024 14:1* **14**, 1–16 (2024).

4. Robles-Martín, A. *et al.* Sub-micro- and nano-sized polyethylene terephthalate deconstruction with engineered protein nanopores. *Nature Catalysis 2023 6:12* **6**, 1174–1185 (2023).

5. Imburgia, C. *et al.* Random access and semantic search in DNA data storage enabled by Cas9 and machine-guided design. *Nature Communications 2025 16:1* **16**, 1–11 (2025).

6. Lopez, R. *et al.* DNA assembly for nanopore data storage readout. *Nature Communications 2019 10:1* **10**, 1–9 (2019).

7. Branton, D. & Deamer, D. *Nanopore Sequencing*. *Nanopore Sequencing* (2019). doi:10.1142/10995.

8. Cai, S., Sze, J. Y. Y., Ivanov, A. P. & Edel, J. B. Small molecule electro-optical binding assay using nanopores. *Nat Commun* **10**, (2019).

9. Cai, S. *et al.* Single-molecule amplification-free multiplexed detection of circulating microRNA cancer biomarkers from serum. *Nat Commun* 1–12 (2021) doi:10.1038/s41467-021-23497-y.

10. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology 2021 39:11* **39**, 1348–1365 (2021).

11. Sze, J. Y. Y., Ivanov, A. P., Cass, A. E. G. & Edel, J. B. Single molecule multiplexed nanopore protein screening in human serum using aptamer modified DNA carriers. *Nat Commun* **8**, 1–10 (2017).

12. Cai, S., Sze, J. Y. Y., Ivanov, A. P. & Edel, J. B. Small molecule electro-optical binding assay using nanopores. *Nat Commun* **10**, (2019).

13. Zhang, X. *et al.* Specific Detection of Proteins by a Nanobody-Functionalized Nanopore Sensor. *ACS Nano* **17**, 9167–9177 (2023).

14. Raveendran, M., Lee, A. J., Sharma, R., Wälti, C. & Actis, P. Rational design of DNA nanostructures for single molecule biosensing. *Nature Communications 2020 11:1* **11**, 1–9 (2020).

15. Bell, N. A. W. & Keyser, U. F. Digitally encoded DNA nanostructures for multiplexed, single-molecule protein sensing with nanopores. *Nature Nanotechnology 2016 11:7* **11**, 645–651 (2016).

16. Ren, R. *et al.* Multiplexed detection of viral antigen and RNA using nanopore sensing and encoded molecular probes. *Nature Communications 2023 14:1* **14**, 1–16 (2023).

17. Koch, C. *et al.* Nanopore sequencing of DNA-barcoded probes for highly multiplexed detection of microRNA, proteins and small biomarkers. *Nature Nanotechnology 2023 18:12* **18**, 1483–1491 (2023).

18. Lecun, Y., Eon Bottou, L., Bengio, Y. & Abstract|, P. H. *Gradient-Based Learning Applied to Document Recognition*.

19. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 (1948).