WEP Warwick Evans Publishing

Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 4
NSCBM 2024

Revolutionizing Cancer Genomics: AI-Driven Interpretation of Nanopore Sequencing Signals

Dan Sun *

Tempus AI, Inc., Chicago, IL, USA

* Corresponding Author Email: dan.sun@tempus.com

Abstract. Nanopore sequencing technology enables the detailed analysis of biomolecules such as DNA and RNA by detecting variations in ionic current as these molecules traverse nano-scale pores. This method captures nuanced patterns, including sequence context and modification signatures. This research investigates the application of artificial intelligence (AI), particularly deep learning models such as SquiggleNet, NanoDeep, and DeepMod2, in analyzing nanopore sequencing data. The performance metrics of these models are notable, with SquiggleNet achieving an accuracy of 90.8% and an AUC score of 0.817, along with a recall rate of 72.5%. NanoDeep demonstrated an AUC score of up to 0.925 on simulated datasets and an accuracy of 84.9%, while DeepMod2 exhibited varying AUC scores with the highest being 0.903 and recall rates reaching 92.9%. The findings underscore the transformative potential of AI on enhancing clinical diagnosis and customizing medical treatment through rapid and precise biomarker detection. However, despite these promising results, further research and validation are necessary to confirm the efficacy and robustness of these AI models in diverse clinical settings. This research highlights the significant advancements that AI can bring to the field of genomics, especially in cancer diagnostics, but emphasizes the need for continued development and real-world testing to fully realize their potential.

Keywords: Nanopore sequencing; deep learning; biomarkers detection; artificial intelligence in genomics; cancer diagnosis.

1. Introduction

Cancer remains as one of the most formidable health challenges globally. In 2024, an estimated 2,001,140 new cancer cases and 611,720 cancer-related deaths are expected to occur in the United States [1], highlighting the widespread impact of cancer and the persistent obstacles in its prevention, identification, and management. Despite the advancements that have been made in medical science to continuously improve the health outcomes and reduce mortality rate, there is an increasing incidence rate for several major cancer types, such as breast, colon, lung and bronchus, and prostate cancers.

In certain types of cancers, early-stage detection is a significant challenge due to the aggressive and asymptomatic pathology, leading to late diagnoses and poor prognoses. For example, ovarian cancer typically remains undetected until it has metastasis as early-stage symptoms are minimal and nonspecific, making effective screening critical to improve outcomes [2]. Similarly, breast cancer, despite effectiveness in screening technologies like mammography, still sees a large number of late-stage diagnoses in under-screened populations, highlighting the need for more accessible screening tools [3]. The usage of computed tomography (CT) and biopsy, as the traditional methods for detecting lung cancer, often leads to limitations such as low sensitivity for small lesions and invasiveness, respectively. Detecting non-small cell lung cancer (NSCLC) markers at an early stage is still a challenge due to the low concentration of tumor markers, causing the insufficiency for early diagnosis [4]. These highlight the urgent need of innovative diagnostic technologies and treatments in cancer care, particularly in the early-stage detection where the potential for successful treatment is the highest.

Biosensing consists of analytical instruments that combine biological sensing components, such as enzymes and antibodies [5], to identify biological or chemical reactions and transform the results into an electronic signal. These devices are crucial from environment monitoring to medical diagnostics due to its capability of detecting and analyzing changes in substances that may indicate the presence of diseases, toxins, or other biological phenomena. Nanopore sensors offer a promising advancement for early-stage cancer diagnostics. Utilizing nanoscale pores, these sensors enable rapid, sensitive, and real-time detection of biomolecules, such as DNA, RNA, proteins, and peptides [6-8]. As molecules pass through or interact with the nanopores, their distinct characteristics generate unique signals that facilitate the determination of the molecules' properties and concentrations. Compared to traditional invasive and time-consuming diagnostic methods like biopsies and imaging, nanopore sensors provide a less invasive and quicker alternative, potentially leading to earlier detection of malignancies.

The data generated by nanopore sensors is both intricate and voluminous, necessitating advanced analytical techniques. Artificial intelligence (AI), specifically machine learning and deep learning, is pivotal in enhancing the analysis of this data. Deep learning models, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have shown promise in improving the accuracy and efficiency of interpreting nanopore data interpretation [9]. The synergy between AI and nanopore sequencing technology has the potential to advance early-stage cancer diagnostics, contributing to the field of precision medicine. This research will analyze the application performance of AI in cancer detection and analysis.

2. Nanopore sensing technology

Nanopore-based sensing technology involves using nanopores—tiny holes that are just a few nanometers in diameter—to analyze biological molecules such as DNA, RNA, miRNA, and proteins. Nanopores can be biological, formed by protein subunits within lipid bilayers or solid-state, created in synthetic membranes [8]. As a biomolecule moves through a nanopore, it changes the flow of ionic current passing through the pore. These changes in ionic current are then detected and recorded as time-dependent current signals. To detect the presence of the analyte effectively, the nanopore dimensions should closely match the size of the analyte, ensuring a measurable change in ionic current that surpasses the noise level [7]. Under an applied potential, the target biomolecule enters the nanopore, causing temporary disruptions or blockades in ionic current. These disruptions vary in amplitude, duration, and frequency, reflecting the unique sequence and structural properties of the translocating biomolecules. It is of great importance to extract and interpret these patterns for cancer studies because it enables the identification of genetic and epigenetic changes that drive cancer development and progression.

3. Traditional signal processing techniques in nanopore data analysis

Accurate interpretation of the raw electrical signals requires sophisticated signal processing techniques to filter out noise and enhance signal clarity. Traditional signal processing techniques have been fundamental in the early stages of nanopore sequencing data analysis. Methods such as Fourier transform, and wavelet transform have been utilized to enhance signal quality and extract meaningful features from the noisy raw data generated during sequencing. Fourier transform allows the conversion of time-domain signals into the frequency domain, aiding in the identification and filtration of noise components, thereby isolating critical signal characteristics essential for accurate sequence analysis. Wavelet transform, effective for non-stationary signals like those from nanopore sequencing, provides a multi-resolution analysis that captures transient features and subtle variations in the data.

However, limitations of those traditional methods such as manual feature extraction, the high dimensionality and inherent noise levels of nanopore sequencing data hinder the overall efficiency, potentially leading to inaccuracies. It is also important to note that traditional techniques may not effectively capture the intricate dependencies and patterns within the signal data, which are paramount for precise genetic and epigenetic analysis. These limitations necessitate more advanced approaches to fully exploit the rich information embedded in nanopore sequencing data.

4. AI in analyzing nanopore sequencing data

Recent advancements in AI and computing power have significantly enhanced the analysis of nanopore sequencing data, addressing many limitations of traditional signal processing techniques. CNNs and Recurrent Neural Networks (RNNs) have shown remarkable capabilities in automatically learning hierarchical representations from raw data. Unlike traditional approaches, these models do not require manual feature extraction and can capture both local and long-range dependencies within the data. CNNs are particularly effective in recognizing spatial patterns, while RNNs, including advanced variants like LSTM networks, excel in modeling temporal dynamics and contextual relationships. The advent of AI has unblocked new possibilities for genomic research and clinical applications, specifically in early diagnosis and personalized treatment strategies for cancer.

4.1. SquiggleNet: Real-time classification of nanopore signals

SquiggleNet is a pioneering deep neural network created to directly categorize nanopore reads based on electrical signals [10], as shown in Fig. 1. It represents a breakthrough with its ability to classify these reads promptly with a 90.8% overall accuracy in the Respiratory Metagenome dataset, 72.5% true positive rate, and 90.9% true negative rate. This level of performance is a notable achievement given the complexity of the task and the constraints of real-time processing. Besides, SquiggleNet operates with minimal memory requirements (304 KB of RAM) and substantially fewer computational resources compared to traditional alignment-based methods that require extensive data processing and large genome indexes.

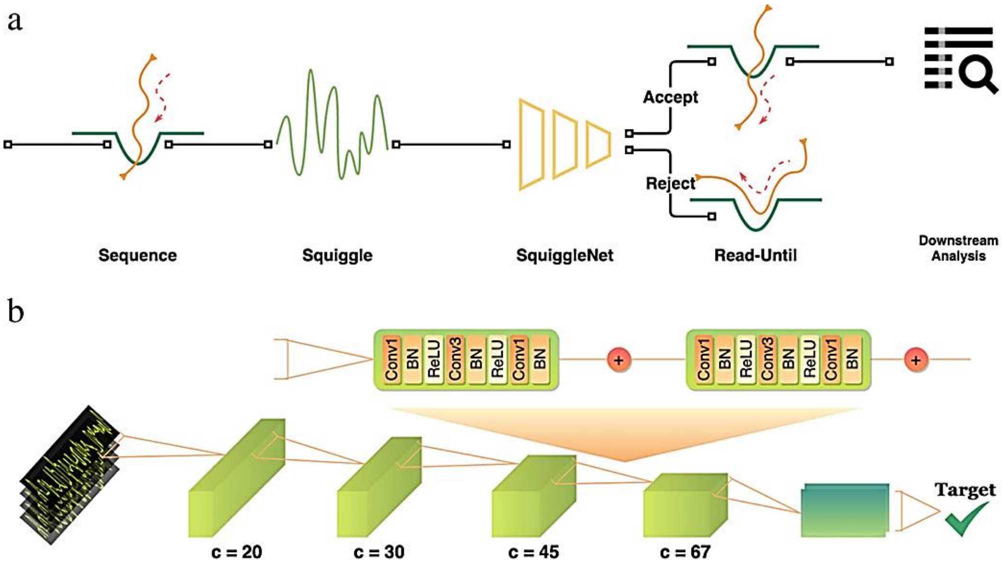

Fig. 1 SquiggleNet-based nanopore analysis. (a) SquiggleNet classifies DNA molecules based on electric signals during nanopore translocation, sequencing accepted molecules to full length and ejecting others. (b) Application of 1D-ResNet-styled bottleneck blocks, average pooling, and a fully connected layer for classification [10].

The network is a binary classifier established using a convolutional architecture with modified residual blocks to processes the nanopore signals generated during DNA translocation. These signals, also known as “squiggles”, are indicative of the nucleotides passing through the nanopores, during which SquiggleNet swiftly analyzes these signals to determine if the molecule is of interest. If the classifier deems the molecule valuable, it continues to sequence the entire length of the DNA. If not, the molecule is ejected, allowing the nanopore to process another molecule. SquiggleNet’s rapid and precise classification capabilities are evidenced by its success in the Respiratory Metagenome dataset, making it a valuable tool for detecting critical DNA modifications such as methylation, essential biomarkers for various diseases, including cancer.

Compared to traditional base-calling and sequence alignment methods that demand extensive CPU/GPU resources, SquiggleNet significantly improves speed and efficiency, enabling accurate classification without the need for large computational infrastructure. Its ability to generalize to unseen bacterial species further underscore its versatility and potential applications in clinical diagnostics and field research. By leveraging advanced deep learning techniques, SquiggleNet sets a new benchmark for nanopore sequencing, facilitating faster, more accurate, and resource-efficient genomic analysis.

4.2. NanoDeep: Advancing real-time genomic analysis

NanoDeep is an innovative deep learning framework designed to improve adaptive sampling in nanopore sequencing [9], as shown in Fig. 2. The model leverages a CNN with squeeze-and-excitation (SE) modules to significantly improve the real-time identification and enrichment of microbial sequences from mixed samples. By analyzing the raw squiggle data, NanoDeep can selectively sequence microbial DNA while depleting human DNA, thereby enhancing sequencing efficiency. Compared to previous models, NanoDeep achieves an accuracy of 0.849 in simulated datasets and 0.752 in real-world applications, with an impressive AUC of 0.925. This level of performance outshines models like DeepSelectNet (AUC: 0.888, Accuracy: 0.804) and SquiggleNet (AUC: 0.867, Accuracy: 0.771) [9]. NanoDeep processes reads at a significantly faster rate (2.89 milliseconds per 50 reads) than both DeepSelectNet and SquiggleNet, emphasizing its suitability for real-world applications where speed is essential.

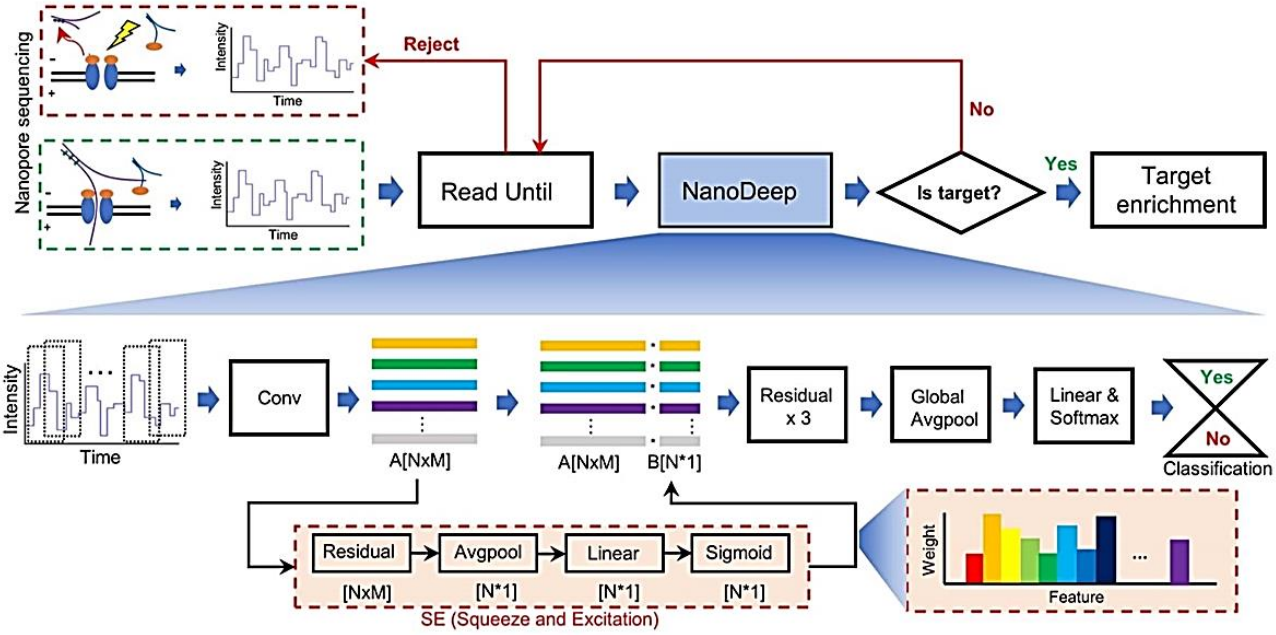

Fig. 2 NanoDeep's framework utilizes the ReadUntil utility for real-time raw signal acquisition and employs a deep neural network to classify signal fragments, rejecting non-target sequences. The network architecture includes a CNN, SE modules, and a Residual network for accurate prediction [9].

NanoDeep's integration of sophisticated AI models has profound implications for a wide range of biomedical research and diagnostic applications, which not only enhances sequencing efficiency but also sets a new benchmark in real-time genomic analysis. For instance, in clinical diagnostics, NanoDeep's ability to enrich microbial sequences while depleting human DNA enhances the accuracy and efficiency of pathogen identification and cancer biomarker detection. This ensures more reliable diagnostic outcomes, which is pivotal for timely and effective patient care. Additionally, the framework's robust performance facilitates advanced research in environmental microbiomes, allowing for precise identification of microbial communities in various ecosystems. This supports a deeper understanding of ecological dynamics and environmental health management. Overall, the AI-driven improvements in speed and accuracy enable early diagnosis and personalized treatment strategies in oncology, leading to better patient prognosis and treatment efficacy.

4.3. DeepMod2: Enhancements in methylation analysis

DeepMod2 is a cutting-edge deep learning framework building upon the capabilities of its predecessor, DeepMod [11, 12], as shown in Fig. 3. It revolutionizes the speed and precision of DNA methylation detection. By harnessing deep learning techniques, the framework implements both a bidirectional long short-term memory (BiLSTM) model and a Transformer model to analyze POD5 and FAST5 signal files, as well as aligned read sequences from BAM files. This comprehensive approach enables DeepMod2 to predict 5mC methylation at single-read and single-site levels. The breakthrough of DeepMod2 lies in its ability to accurately and efficiently analyze methylation data from the latest Oxford Nanopore sequencing data formats (POD5/FAST5) and flowcells (R10.4) using advanced deep learning architectures. This sets DeepMod2 apart from previous detection tools, which were limited to older formats and less sophisticated models.

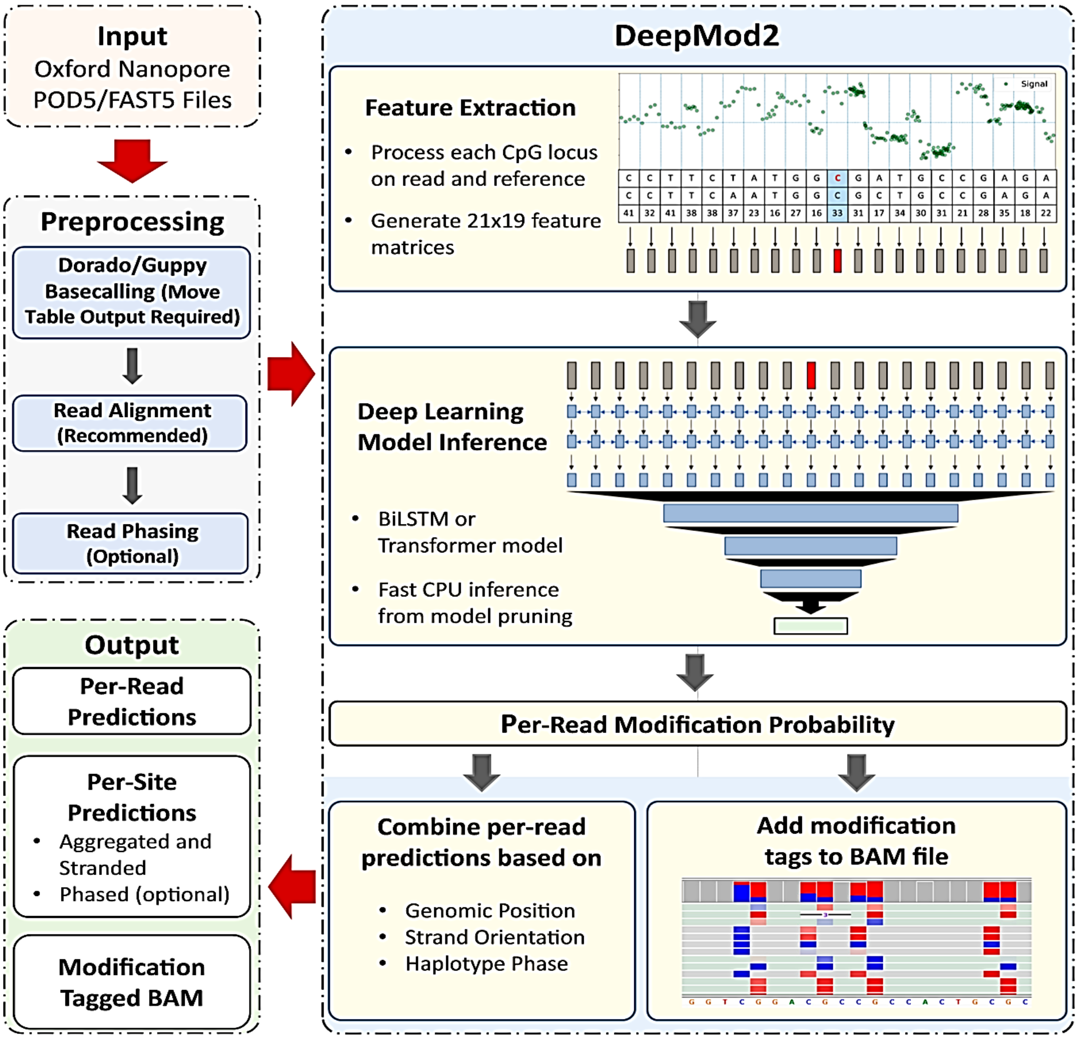
Input
Oxford Nanopore POD5/FAST5 Files
Preprocessing
Dorado/Guppy Basecalling (Move Table Output Required)
Read Alignment (Recommended)
Read Phasing (Optional)
Output
Per-Read Predictions
Per-Site Predictions
• Aggregated and Stranded
• Phased (optional)
Modification Tagged BAM
DeepMod2
Feature Extraction
• Process each CpG locus on read and reference
• Generate 21x19 feature matrices
Signal
C C T T C T A T G G C G A T G C C G A G A
C C T T C A A T G G C G C T G C C G A G A
41 32 41 38 38 37 23 16 27 16 33 31 17 34 30 31 21 28 35 18 22
Deep Learning Model Inference
• BiLSTM or Transformer model
• Fast CPU inference from model pruning
Per-Read Modification Probability
Combine per-read predictions based on
• Genomic Position
• Strand Orientation
• Haplotype Phase
Add modification tags to BAM file
3
G G T C G G A C G C C G C C A C T G C G C

Fig. 3 DeepMod2 processes POD5 or FAST5 files and a BAM file with read sequences for methylation calling. It utilizes a feature extraction module and a BiLSTM or Transformer model to make per-read predictions, outputting both aggregated and stranded methylation data [12].

Benchmark evaluations indicate that DeepMod2 outperforms other state-of-the-art methylation callers, with F1-scores ranging from 95.7% to 98.2% achieved on R9.4.1 and R10.4.1 flowcell datasets, respectively. Precision rates reach as high as 99.8%, with recall rates around 98.5%, highlighting its accuracy in detecting methylation sites. DeepMod2 optimizes computational efficiency by employing pruning techniques to reduce the computational load of large BiLSTM layers, making it ideal for high-throughput sequencing applications. The model's unique ability to process R9.4 and R10.4 flowcell data further broadens its applicability, ensuring seamless integration into diverse genomic research workflows. This versatility underscores DeepMod2's potential for accurate methylation mapping, which is crucial for cancer biomarker discovery and therapeutic targeting.

5. Conclusion

This research underscores the substantial potential of integrating AI, especially deep learning models, with nanopore sequencing technology for cancer diagnostics. Models such as SquiggleNet, NanoDeep, and DeepMod2 exhibit exceptional accuracy and high recall rates, ensuring the reliable identification of critical genetic and epigenetic markers. These advancements suggest a significant potential to revolutionize clinical diagnostics and personalized medicine in oncology. Despite these promising results, further validation in diverse clinical settings is necessary to ensure the robustness and generalizability of these models. Future research should focus on optimizing these models for broader applications and developing scalable, user-friendly tools for clinical adoption. Addressing these challenges is crucial for translating technological advancements into clinical benefits, ultimately improving cancer patient outcomes. The fusion of AI with nanopore sequencing paves the way for earlier and more precise cancer diagnostics, advancing the field of precision oncology. By enhancing the accuracy and efficiency of biomarker detection, this integration holds promise for significantly improving patient prognosis and treatment efficacy in oncology.

References

[1] Siegel R L, Giaquinto A N, Jemal A. Cancer statistics, 2024. CA: a cancer journal for clinicians, 2024, 74(1): 12-49.

[2] Nash Z, Menon U. Ovarian cancer screening: Current status and future directions. Best practice & research Clinical obstetrics & gynaecology, 2020, 65: 32-45.

[3] Smith R A, Andrews K S, Brooks D, et al. Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening. CA: a cancer journal for clinicians, 2019, 69(3): 184-210.

[4] Chen S, Li M, Weng T, et al. Recent Progresses of Biosensors for the Detection of Lung Cancer Markers. Journal of Materials Chemistry B, 2023,11: 5715-5747.

[5] Singh A, Sharma A, Ahmed A, et al. Recent advances in electrochemical biosensors: Applications, challenges, and future scope. Biosensors, 2021, 11(9): 336.

[6] Wang Y, Zhao Y, Bollas A, et al. Nanopore sequencing technology, bioinformatics and applications. Nature biotechnology, 2021, 39(11): 1348-1365.

[7] Ying Y L, Hu Z L, Zhang S, et al. Nanopore-based technologies beyond DNA sequencing. Nature nanotechnology, 2022, 17(11): 1136-1146.

[8] Kim H J, Choi U J, Kim H, et al. Translocation of DNA and protein through a sequentially polymerized polyurea nanopore. Nanoscale, 2019, 11(2): 444-453.

[9] Lin Y, Zhang Y, Sun H, et al. NanoDeep: a deep learning framework for nanopore adaptive sampling on microbial sequencing. Briefings in Bioinformatics, 2024, 25(1): bbad499.

[10] Bao Y, Wadden J, Erb-Downward J R, et al. SquiggleNet: real-time, direct classification of nanopore signals. Genome biology, 2021, 22: 1-16.

[11] Liu Q, Fang L, Yu G, et al. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. Nature communications, 2019, 10(1): 2449.

[12] Ahsan M U, Gouru A, Chan J, et al. A signal processing and deep learning framework for methylation detection using Oxford Nanopore sequencing. Nature Communications, 2024, 15(1): 1448.