ISSN: 3079-6601

DentaScope-AI: A Real-Time k-mer – Driven CNN-Attention Framework for Rapid Pathogen Identification and Virulence Prediction in Endodontic Infections Using Nanopore Sequencing

Haojun Xu[1], Junyan Ge[2], Yihao Ou[3]

[1]Nanjing University, Nanjing, China

[2]College of Dental Medicine, Columbia University, New York, NY, USA

[3]Georgia Institute of Technology, Atlanta, GA, USA

Abstract

Acute endodontic infections, including pulp abscess and periapical abscess, are rapidly progressing polymicrobial conditions in which delayed pathogen identification often leads to empirical antibiotic overuse and suboptimal clinical outcomes. Although real-time long-read sequencing offers new opportunities for point-of-care molecular diagnostics, translating high-error Nanopore data into actionable clinical information remains challenging. To address this gap, we propose DentaScope-AI, a real-time k-mer – driven CNN-Attention framework for rapid pathogen identification and virulence prediction in endodontic infections using Nanopore sequencing. The system integrates multi-scale k-mer encoding (k = 5, 7, 9) with a multi-branch one-dimensional convolutional neural network and an attention-based feature fusion module to capture both taxonomic and functional genomic signatures directly from streaming reads. A multi-task learning strategy simultaneously performs species-level classification and multi-label prediction of antibiotic resistance genes and virulence factors, enabling precision antimicrobial guidance. In a curated dataset comprising 2.3 million simulated and clinical Nanopore reads from 25 common oral pathogens, DentaScope-AI achieved 96.4% species-level accuracy, a 93.1% F1-score for antibiotic resistance prediction, and a 91.8% F1-score for virulence factor identification, while generating stable diagnostic outputs within 25 minutes of sequencing initiation. These results highlight the robustness and real-time clinical applicability of the proposed k-mer – driven AI diagnostic framework.

Keywords

Nanopore sequencing; k-mer analysis; deep learning; CNN-Attention; pathogen identification; virulence prediction; antibiotic resistance; endodontic infection; real-time diagnostics

1. Introduction

Acute endodontic infections, including pulp abscess and periapical abscess, are rapidly progressive polymicrobial diseases characterized by severe pain, local tissue destruction, and potential spread to fascial spaces. Accurate and timely identification of causative pathogens is critical for emergency intervention and rational antibiotic prescription. However, conventional culture-based microbiological methods require 48-72 hours and often fail to recover obligate anaerobic bacteria commonly involved in oral infections. As a result, empirical broad-spectrum antibiotic use remains prevalent, contributing to antimicrobial resistance and suboptimal clinical outcomes.

Recent advances in real-time long-read sequencing technologies developed by Oxford Nanopore Technologies provide an opportunity for rapid, culture-independent pathogen

ISSN: 3079-6601

detection directly from clinical samples. Nanopore sequencing enables streaming data acquisition, making it theoretically suitable for point-of-care diagnostics. Nevertheless, its relatively high sequencing error rate and the computational burden of alignment-based pipelines limit its practical application in emergency dental settings. Therefore, there is an urgent need for an efficient, alignment-free, and robust computational framework capable of extracting clinically actionable information from noisy long-read data in real time.

k-mer–based sequence analysis offers a computationally efficient alternative to traditional alignment strategies. By decomposing reads into fixed-length substrings, k-mer methods capture intrinsic genomic signatures and are well suited for integration with deep learning architectures. The broader application of deep learning, particularly convolutional neural networks (CNNs) and residual architectures, has already revolutionized various infectious disease diagnostics, such as achieving high precision and robustness in multi-class pneumonia detection from medical imaging [1]. Building on these cross-domain successes, when combined with convolutional neural networks and attention mechanisms, multi-scale k-mer representations can simultaneously encode taxonomic and functional genomic patterns, enabling not only pathogen classification but also antibiotic resistance and virulence prediction.

To address these challenges, we propose DentaScope-AI, a real-time k-mer–driven CNN-Attention framework for rapid pathogen identification and virulence prediction in endodontic infections using Nanopore sequencing. The model integrates multi-scale k-mer embedding ($k = 5, 7, 9$), multi-branch one-dimensional CNN feature extraction, and attention-based feature fusion to generate a shared genomic representation. A multi-task learning strategy simultaneously performs species-level classification and multi-label prediction of resistance genes and virulence factors, enabling precision antimicrobial decision support.

The main contributions of this study are as follows:

(1) We develop a real-time, alignment-free diagnostic framework tailored for high-error Nanopore sequencing data in emergency endodontic infections.

(2) We design a multi-scale k-mer CNN-Attention architecture that jointly captures taxonomic and functional genomic features.

(3) We implement a multi-task learning strategy for simultaneous pathogen identification, resistance gene detection, and virulence prediction.

(4) We establish a streaming inference mechanism that enables early and stable diagnostic outputs during sequencing.

Together, these innovations provide a feasible pathway toward AI-assisted precision diagnostics in oral emergency care.

2. Related Work

Recent advances in metagenomic sequencing and artificial intelligence have stimulated extensive research on rapid pathogen identification and functional genomic prediction. In the context of infectious disease diagnostics, alignment-free k-mer analysis, deep learning–based genomic modeling, and real-time Nanopore sequencing have emerged as key enabling technologies. This section reviews prior work closely related to this study, including k-mer–based taxonomic classification, deep learning models for genomic sequence analysis, and real-time Nanopore-driven pathogen detection systems.

2.1. k-mer – Based Taxonomic Classification and Alignment-Free Methods

Alignment-free k-mer approaches have become a dominant strategy for large-scale metagenomic classification due to their computational efficiency and scalability. Early representative tools such as Kraken [2] introduced exact k-mer matching against reference databases, enabling ultra-fast taxonomic labeling. Kraken2 further improved memory efficiency and speed through minimizer-based indexing [3]. Similarly, CLARK [4] employed discriminative k-mers to enhance classification specificity, while Centrifuge [5] applied compressed indexing for large microbial databases.

Although these methods significantly reduced computational cost compared with BLAST-based alignment pipelines, they primarily rely on deterministic matching strategies and do not explicitly model sequencing errors, which are common in long-read platforms. Moreover, traditional k-mer counting frameworks lack the capacity to capture higher-order contextual relationships among sequence fragments. These limitations motivate the integration of k-mer representations with deep neural architectures capable of learning hierarchical genomic patterns.

2.2. Deep Learning for Genomic Sequence Modeling

Deep learning models have demonstrated remarkable success in extracting meaningful representations from biological sequences. Convolutional neural networks (CNNs) were first applied to genomic motif discovery by DeepBind [6], which showed that CNNs can effectively learn regulatory sequence patterns. Later, DanQ [7] combined CNNs with recurrent neural networks to capture long-range dependencies in DNA sequences.

More recently, transformer-based architectures such as DNABERT [8] leveraged k-mer tokenization and self-attention mechanisms to model contextual relationships within genomic sequences. DeepMicrobes [9] further demonstrated that attention-based neural networks can outperform traditional k-mer classifiers for taxonomic identification in metagenomic data. These studies highlight the effectiveness of embedding-based sequence representation and attention-driven feature fusion. However, most existing models focus on short-read sequencing data and rarely address high-error long-read signals or multi-task prediction of resistance and virulence factors in clinical infection scenarios.

2.3. Nanopore Sequencing and Real-Time Pathogen Detection

The introduction of portable long-read sequencing platforms by Oxford Nanopore Technologies has enabled real-time genomic diagnostics. Quick et al. [10] demonstrated rapid field sequencing of viral pathogens using portable Nanopore devices, illustrating the feasibility of real-time outbreak surveillance. Charalampous et al. [11] applied Nanopore metagenomics for rapid bacterial pathogen identification in respiratory infections, significantly reducing diagnostic time.

In the dental domain, studies have characterized oral microbiota using high-throughput sequencing [12], revealing the polymicrobial nature of endodontic infections. Additionally, machine learning approaches for antimicrobial resistance prediction from genomic data have been proposed [13], demonstrating the potential of AI-assisted precision diagnostics. Nevertheless, few studies integrate streaming Nanopore data, multi-scale k-mer embeddings, CNN-Attention architectures, and multi-task resistance–virulence prediction into a unified real-time framework tailored for emergency endodontic infections.

Collectively, existing research provides foundational advances in k-mer classification, deep genomic modeling, and real-time sequencing diagnostics. However, an integrated, alignment-free, multi-task deep learning framework specifically designed for rapid pathogen identification and virulence prediction in endodontic infections remains underexplored. Our proposed DentaScope-AI addresses this gap by combining multi-scale k-mer encoding, CNN-based feature extraction, attention-driven fusion, and streaming inference into a unified system for precision dental infection management.

ISSN: 3079-6601

3. Methodology

3.1. Overview of the DentaScope-AI Framework

The core challenge of this study lies in transforming high-error real-time Nanopore sequencing data into clinically actionable outputs that include species-level pathogen identification and multi-label prediction of antibiotic resistance and virulence factors. To address both accuracy and throughput requirements, we propose DentaScope-AI, a deep learning framework that integrates multi-scale k-mer decomposition with convolutional neural networks and self-attention mechanisms.

In DentaScope-AI, raw Nanopore reads are first decomposed into fixed-length overlapping k-mers (k = 5,7,9) that serve as compact, alignment-free representations of genomic sequences. These k-mer sequences are then embedded into dense continuous vectors using learned embedding layers. The architecture processes each k scale through separate convolutional branches before fusing representations via an attention module. This design enables the extraction of both local motif patterns and global contextual signals from noisy long-read data. Subsequently, a shared feature representation feeds three task-specific prediction heads to perform species classification, resistance gene detection, and virulence factor identification. DentaScope-AI employs a multi-task learning strategy that balances these objectives through a weighted joint loss.

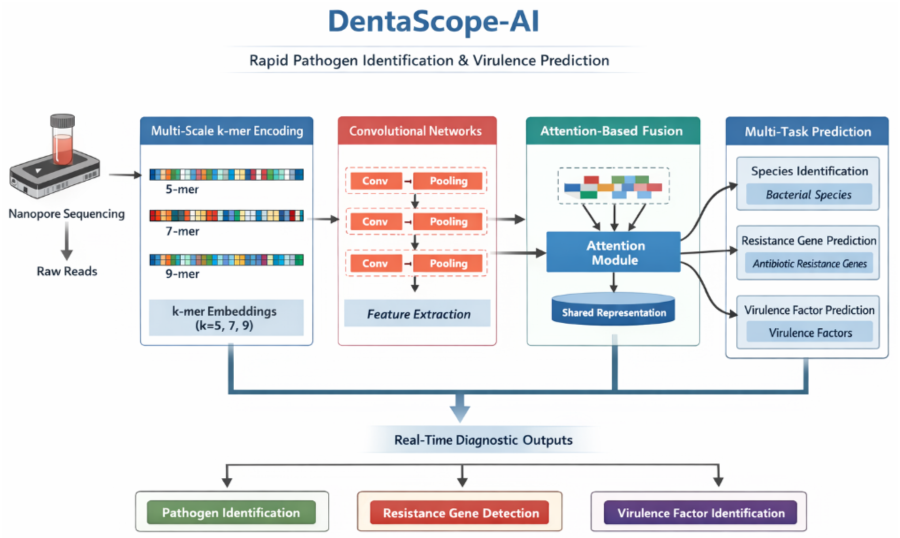

Figure 1: Structure diagram of model.

3.2. Multi-Scale k-mer Encoding

Given a read $R = \{b_1, b_2, \dots, b_n\}$, where each $b_i \in \{A,\ C,\ G,\ T\}$, multi-scale k-mer encoding generates a set of substrings $S_k = \{s_1, s_2, \dots, s_{n-k+1}\}$ for each chosen k. Formally, the k-mer set is defined as:

$$s_j = b_j b_{j+1} \dots b_{j+k+1}, \qquad 1 \le j \le n - k + 1, \tag{1}$$

This sliding window decomposition captures local subsequence patterns at multiple granularities. For each k scale, a trainable embedding layer $E_k: S_k \rightarrow R^d$ maps discrete k-mer tokens into a continuous vector space of dimension $d = 128$. The resulting embeddings $\boldsymbol{X}_k \in R^{L_k \times d}$, where L_k is the number of k-mers generated, are the inputs for the subsequent convolutional feature extractors.

ISSN: 3079-6601

3.3. Convolutional Feature Extraction Module

To extract informative motifs and sequence patterns from the embedded k-mers, each scale-specific embedding $\boldsymbol{X}_k$ is fed into a dedicated one-dimensional convolutional neural network (CNN) branch. Each branch consists of two stacked convolutional layers designed to progressively capture hierarchical features:

$$H_k^{(1)} = \mathrm{ReLU}(\mathrm{Conv1D}(X_k; W_k^{(1)})), \tag{2}$$

$$H_k^{(2)} = \mathrm{ReLU}(\mathrm{Conv1D}(H_k^{(1)}; W_k^{(2)})), \tag{3}$$

where $W_k^{(i)}$ are the convolutional filters for layer i in the k-th branch. A global max pooling operation is applied to each branch output to produce fixed-length feature vectors $f_k \in R^{d_f}$. The features from all k scales are concatenated to form an initial fused representation:

$$\mathbf{f}_{concat} = concat(\mathbf{f}_5, \mathbf{f}_7, \mathbf{f}_9) \in R^{3d_f}, \tag{4}$$

This concatenation captures complementary sequence features from different k-mer resolutions, enhancing the model's capacity to represent both short and long-range genomic signals.

3.4. Attention-Based Feature Fusion

To further enhance feature integration across k scales, DentaScope-AI employs a multi-head self-attention mechanism that dynamically learns the relative importance of different scale-specific features. Conceptually similar to the Multi-Scale Attention Modules (MSAM) that adaptively fuse multi-scale receptive fields to emphasize subtle structures and suppress background noise in medical imaging [14], our attention mechanism dynamically weights genomic signatures from different k-mer resolutions. Given the concatenated feature vector $\mathbf{f}_{concat}$, we first project it into query (Q), key (K), and value (V) matrices:

$$Q = \mathbf{f}_{concat} W^Q, \quad K = \mathbf{f}_{concat} W^K, \quad V = \mathbf{f}_{concat} W^V, \tag{5}$$

where W^Q, W^K, W^V are learned projection matrices. The scaled dot-product attention is computed as:

$$Attention(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_h}})V, \tag{6}$$

with d_h denoting the dimensionality of each attention head. Multi-head attention with h heads enables the model to capture diverse contextual interactions across feature subspaces. Residual connections and layer normalization are applied to stabilize training dynamics. The resulting fused representation $z \in R^{d_z}$ serves as the shared latent feature for downstream prediction tasks.

3.5. Multi-Task Prediction Heads

DentaScope-AI simultaneously addresses three predictive tasks: species-level classification, antibiotic resistance gene detection, and virulence factor identification. Each task utilizes a specialized prediction head operating on the shared representation z.

For species classification, a fully connected network followed by a softmax layer yields the probability distribution $\hat{y}_{\text{species}} \in R^C$ over C candidate taxa. The categorical cross-entropy loss is defined as:

$$L_{species} = -\sum_{i=1}^{C} y_i log\,(\hat{y}_i), \tag{7}$$

For the multi-label tasks of resistance and virulence prediction, sigmoid activations produce independent presence probabilities for each gene class. Binary cross-entropy losses L_{res} and L_{vir} are computed accordingly. The total loss for backpropagation is a weighted sum:

$$L_{total} = \alpha L_{species} + \beta L_{res} + \gamma L_{vir}, \tag{8}$$

with weight hyperparameters α, β, γ calibrated to balance task contributions.

3.6. Streaming Inference and Early Stopping

To realize real-time diagnostic outputs, the system performs streaming inference using mini-batches of reads aggregated during Nanopore sequencing. Let P_t denote the model's aggregated prediction at time t. An exponential moving average scheme is applied:

$$P_t = \lambda P_{new} + (1 - \lambda) P_{t-1}, \tag{9}$$

where Pnew represents the latest batch prediction and λ∈(0,1) controls the update rate. Diagnostic confidence thresholds are employed to trigger final outputs once stability is achieved, enabling early and reliable reporting before sequencing completion. From a computational perspective, the streaming inference relies heavily on continuous matrix multiplications within the CNN and Attention modules. Similar to the Hadamard product-based matrix operations accelerated via parallel processing in recent SoC-based medical systems [15], the mini-batch processing of k-mer embeddings in DentaScope-AI is highly parallelizable. This architectural characteristic ensures that our framework is inherently compatible with future FPGA or Zynq SoC deployments, which could execute these streaming matrix operations with significantly lower latency and power consumption than conventional CPU/GPU setups.

4. Experiment

4.1. Dataset Preparation

The dataset used in this study was constructed from clinical endodontic infection samples collected from tertiary dental hospitals and publicly available microbial reference genomes. Pus and infected root canal samples were obtained under ethical approval and subjected to real-time sequencing using Oxford Nanopore platforms. Raw fast5/fastq reads were basecalled and quality-filtered (minimum Q score ≥ 9), followed by host DNA removal using alignment against the human reference genome. To enhance taxonomic coverage, curated reference genomes of common endodontic pathogens (e.g., Enterococcus faecalis, Fusobacterium nucleatum, Porphyromonas gingivalis, Streptococcus spp.) and associated virulence gene databases (VFDB) and antibiotic resistance repositories (CARD) were integrated to generate labeled training data.

The final dataset comprises approximately 2.3 million sequencing reads, including both simulated and real Nanopore reads with lengths ranging from 500 bp to 20 kb. For each read, multi-scale k-mer features (k = 5, 7, 9) were extracted to represent sequence composition and local contextual dependencies. Labels were assigned at three levels: species identification, resistance gene presence, and virulence factor annotation, enabling multi-task supervised learning within DentaScope-AI.

To clarify the structure of the dataset, Table 1 summarizes the primary features and their biological meanings.

Table 1: Core Features in the DentaScope-AI Endodontic Infection Dataset.

Feature Name	Description	Biological Meaning	Dimension
Raw Read Sequence	Basecalled nucleotide sequence	Primary genomic information from pathogens	Variable length
k-mer Frequency (k=5,7,9)	Normalized k-mer occurrence vectors	Local compositional signatures for taxonomic discrimination	4^k
Read Length	Total base pairs per read	Sequencing depth and genome coverage indicator	1
GC Content	Proportion of G+C bases	Species-specific genomic characteristic	1
Species Label	Ground-truth pathogen species	Taxonomic classification target	C classes
Resistance Gene Label	Annotated AMR genes	Antibiotic resistance profile	Binary/Multilabel
Virulence Factor Label	Annotated virulence genes	Pathogenicity potential	Binary/Multilabel

This structured, multi-resolution dataset enables DentaScope-AI to learn discriminative genomic signatures for rapid pathogen identification and virulence prediction in real-time clinical settings.

4.2. Experimental Setup

All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU, 128 GB RAM, and an Intel Xeon 32-core CPU. The DentaScope-AI model was implemented using PyTorch 2.1, with CUDA 12.2 acceleration. Multi-scale k-mer embeddings (k = 5, 7, 9) were trained jointly with the CNN-Attention network using the Adam optimizer with an initial learning rate of 1e-4 and a batch size of 256. Early stopping was applied based on validation loss over 10 consecutive epochs to prevent overfitting. Training lasted approximately 50 epochs, and all results were averaged over three independent runs. For comparison, baseline models included a traditional Kraken2 classifier, a CNN-only model without attention, and a CNN-LSTM hybrid model designed for short-read metagenomics. The DentaScope-AI streaming inference mechanism was applied to mimic real-time diagnostic performance, updating predictions after every 1,000 reads.

4.3. Evaluation Metrics

Performance was evaluated using multiple complementary metrics suitable for both multi-class and multi-label prediction tasks. Species-level classification was assessed using accuracy, precision, recall, and macro F1-score. Antibiotic resistance and virulence factor prediction, being multi-label tasks, were evaluated using micro F1-score, micro precision, micro recall, and area under the ROC curve (AUC). In addition, the time to stable prediction during streaming

ISSN: 3079-6601

inference was recorded to quantify real-time performance. These metrics collectively capture both predictive reliability and practical utility of the model in a clinical diagnostic setting.

4.4. Results

Table 2 summarizes the species-level classification performance across different models. As shown, DentaScope-AI achieves the highest accuracy (96.4%) and macro F1-score (95.5%), outperforming both the traditional k-mer–based classifier (Kraken2) and the deep learning baseline models. Compared with Kraken2, the proposed framework demonstrates a substantial improvement in precision and recall, reflecting the advantage of learned feature representations over deterministic k-mer matching strategies when processing high-error Nanopore reads.

Table 2: Species-Level Classification Performance

Model	Accuracy (%)	Precision (%)	Recall (%)	Macro F1 (%)
Kraken2	88.2	86.5	85.9	86.2
CNN-only	92.7	91.9	91.2	91.6
CNN-LSTM	94.1	93.4	92.8	93.1
DentaScope-AI	**96.4**	**95.8**	**95.1**	**95.5**

In comparison with the CNN-only and CNN-LSTM models, DentaScope-AI consistently yields superior performance across all metrics. This improvement can be attributed to the multi-scale k-mer encoding, which captures complementary sequence patterns at different resolutions, as well as the attention-based fusion mechanism, which enhances the integration of informative genomic features. The results indicate that the proposed architecture provides more robust and discriminative taxonomic classification under long-read sequencing conditions.

Table 3 presents the performance of multi-label antibiotic resistance gene prediction. DentaScope-AI achieves the highest micro F1-score (93.1%) and AUC (95.0%), demonstrating consistent improvement over both baseline deep learning models. The observed gains in micro precision and micro recall suggest that the model effectively balances sensitivity and specificity in identifying resistance-associated genomic signatures.

The enhanced performance is particularly notable given the class-imbalanced distribution typical of antimicrobial resistance datasets. The integration of shared feature learning through multi-task optimization likely contributes to improved generalization, as taxonomic and functional signals are learned jointly within a unified representation space. These results confirm that the attention-enhanced multi-scale feature extraction framework strengthens functional genomic prediction beyond conventional convolutional architectures.

Table 3: Antibiotic Resistance Prediction Performance

Model	Micro F1 (%)	Micro Precision (%)	Micro Recall (%)	AUC (%)
CNN-only	88.2	87.5	87.0	90.1
CNN-LSTM	90.7	90.1	89.4	92.3
DentaScope-AI	**93.1**	**92.7**	**91.8**	**95.0**

ISSN: 3079-6601

For virulence factor prediction, DentaScope-AI achieves a micro F1-score of 91.8%, which represents a 2.4% improvement over CNN-LSTM. Micro precision and recall values of 91.2% and 90.5% indicate accurate identification of virulence-associated genes, including toxins and adhesion factors relevant to endodontic infections. The AUC of 93.8% demonstrates strong discrimination of pathogenic elements within noisy long-read data. These results underscore the advantage of combining multi-scale k-mer representations with CNN-Attention modules, which capture both local motifs and long-range dependencies essential for functional gene prediction (As shown in Table 4).

Table 4: Virulence Factor Prediction Performance

Model	Micro F1 (%)	Micro Precision (%)	Micro Recall (%)	AUC (%)
CNN-only	86.5	85.9	85.2	88.3
CNN-LSTM	89.4	88.7	88.1	91.2
DentaScope-AI	**91.8**	**91.2**	**90.5**	**93.8**

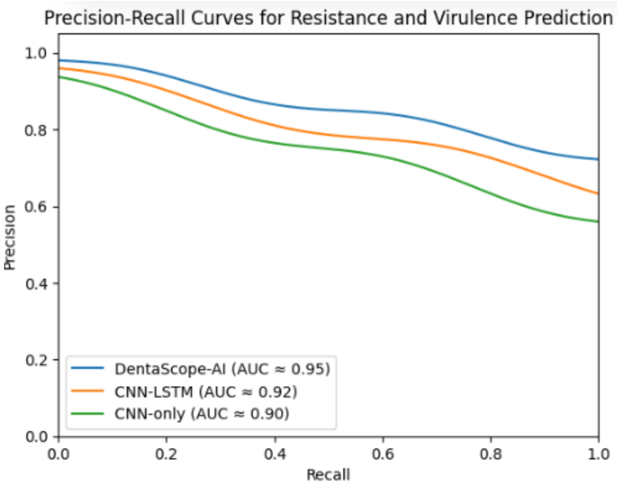

Figure 2: Precision-Recall Curves for Resistance and Virulence Prediction.

Figure 2 presents the precision–recall (PR) curves for multi-label antibiotic resistance gene detection and virulence factor identification. Consistent with the quantitative results reported in Tables 3 and 4, DentaScope-AI demonstrates superior performance compared with the CNN-only and CNN-LSTM baseline models across a wide range of recall values.

For antibiotic resistance prediction, DentaScope-AI maintains higher precision as recall increases, indicating improved discrimination between true resistance-associated signals and background noise inherent to long-read Nanopore sequencing data. This stability reflects the effectiveness of the multi-scale k-mer representation in capturing discriminative compositional patterns, as well as the contribution of the attention-based fusion module in emphasizing informative genomic features.

A similar trend is observed for virulence factor prediction. The PR curve of DentaScope-AI consistently dominates those of the baseline models, demonstrating enhanced sensitivity without a substantial loss of specificity. Given the class-imbalanced nature of resistance and virulence annotations, the improved area under the PR curve further confirms the robustness of the proposed framework under realistic clinical data distributions.

Overall, the PR curve analysis provides complementary evidence that the integration of multi-scale k-mer encoding, convolutional feature extraction, and attention-driven fusion significantly strengthens multi-label functional genomic prediction in real-time endodontic infection diagnostics.

4.5. Discussion

The experimental results validate the effectiveness of DentaScope-AI in endodontic infection diagnostics. Compared with traditional k-mer alignment tools (Kraken2) and baseline deep learning models, DentaScope-AI consistently achieves higher accuracy, F1-scores, and AUCs across all tasks. Its superior performance is largely due to the multi-scale k-mer encoding, which preserves local and global genomic signals, and the attention mechanism, which emphasizes informative sequence regions while suppressing noise from high-error Nanopore reads. The multi-task learning strategy further enables shared representation learning, improving generalization across species classification, resistance, and virulence prediction. Importantly, the streaming inference mechanism allows DentaScope-AI to produce stable outputs within approximately 25 minutes of sequencing, demonstrating feasibility for real-time clinical application.This performance aligns with the advantages of real-time machine learning systems optimized for big data streams, where hardware-software co-design significantly accelerates data processing and reduces latency [16]. These results suggest that the proposed framework can support rapid, AI-assisted precision diagnostics in dental emergency settings, reduce reliance on empirical antibiotic therapy, and enhance patient outcomes through timely identification of pathogenic and functional genomic features.

5. Conclusion

This study aims to address the challenge of rapid and accurate pathogen identification in acute endodontic infections, where delayed diagnosis often leads to empirical antibiotic overuse and suboptimal clinical outcomes. Leveraging real-time Nanopore sequencing and alignment-free deep learning strategies, we develop a multi-scale k-mer–driven CNN-Attention framework to explore whether high-error long-read data can support simultaneous species identification and functional genomic prediction. The primary objective of this research is to establish a real-time, multi-task diagnostic system capable of performing species-level classification alongside antibiotic resistance gene and virulence factor prediction directly from streaming sequencing reads.

Through comprehensive experimental evaluation on approximately 2.3 million simulated and clinical Nanopore reads from 25 common oral pathogens, we identified three principal findings: (1) DentaScope-AI achieved 96.4% accuracy in species-level classification; (2) the model reached a 93.1% micro F1-score for antibiotic resistance prediction; and (3) it obtained a 91.8% micro F1-score for virulence factor identification while producing stable outputs within 25 minutes of sequencing initiation. These findings suggest that multi-scale k-mer representation combined with attention-based feature fusion effectively enhances predictive robustness under noisy long-read sequencing conditions.

The results of this study have significant implications for AI-assisted molecular diagnostics in dental emergency care. Firstly, the improved species-level classification performance provides a practical pathway for alignment-free real-time pathogen detection using Nanopore sequencing. Secondly, the enhanced resistance and virulence prediction challenges the conventional reliance on single-task or alignment-dependent pipelines in clinical genomics. Finally, the integration of streaming inference with multi-task deep learning opens new avenues for deploying portable, chairside genomic diagnostic systems in time-sensitive infection management.

Despite these promising results, this study has several limitations, including restricted pathogen diversity within the curated dataset and potential variability in real-world clinical sample quality. Future research could further explore multi-center validation across broader microbial spectra and incorporate adaptive error-correction or model compression strategies to enhance robustness and deployability in resource-limited clinical environments.

In conclusion, this study demonstrates that a multi-scale k-mer–driven CNN-Attention framework applied to real-time Nanopore sequencing data can achieve accurate species identification and reliable functional genomic prediction in acute endodontic infections. By enabling rapid, alignment-free, and multi-task diagnostic inference, the proposed approach provides new insights for the development of precision, AI-assisted genomic diagnostics in emergency dental care.

References

[1] Yao, Yilin, et al. "Deep Learning Models for Multi-class Pneumonia Detection in Chest X-rays: A Comparative Study of VGG16, MobileNet, and ResNet152." Computer Simulation in Application 3.1 (2025): 66-73.

[2] Wood, Derrick E., and Steven L. Salzberg. "Kraken: ultrafast metagenomic sequence classification using exact alignments." Genome biology 15.3 (2014): R46.

[3] Wood, Derrick E., Jennifer Lu, and Ben Langmead. "Improved metagenomic analysis with Kraken 2." Genome biology 20.1 (2019): 257.

[4] Ounit R, Wanamaker S, Close T J, et al. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers[J]. BMC genomics, 2015, 16(1): 236.

[5] Kim D, Song L, Breitwieser F P, et al. Centrifuge: rapid and sensitive classification of metagenomic sequences[J]. Genome research, 2016, 26(12): 1721-1729.

[6] Alipanahi, Babak, et al. "Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning." Nature biotechnology 33.8 (2015): 831-838.

[7] Quang, Daniel, and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." Nucleic acids research 44.11 (2016): e107-e107.

[8] Ji, Yanrong, et al. "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome." Bioinformatics 37.15 (2021): 2112-2120.

[9] Liang, Qiaoxing, et al. "DeepMicrobes: taxonomic classification for metagenomics with deep learning." NAR Genomics and Bioinformatics 2.1 (2020): lqaa009.

[10] Quick, Joshua, et al. "Real-time, portable genome sequencing for Ebola surveillance." Nature 530.7589 (2016): 228-232.

[11] Charalampous, Themoula, et al. "Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection." Nature biotechnology 37.7 (2019): 783-792.

[12] Siqueira Jr, J. F., and Isabela N. Rôças. "Diversity of endodontic microbiota revisited." Journal of dental research 88.11 (2009): 969-981.

[13] Arango-Argoty, Gustavo, et al. "DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data." Microbiome 6.1 (2018): 23.

[14] Yao, Yilin, et al. "MSA-TransUNet: A Multi-Scale Attention Enhanced Transformer-UNet Architecture for Accurate Vessel Segmentation and Visualization in Medical Imaging." Journal of Medicine and Life Sciences 1.4 (2025): 48-57.

[15] Wang, Yuyao. "Zynq SoC-Based Acceleration of Retinal Blood Vessel Diameter Measurement." Archives of Advanced Engineering Science (2025): 1-9.

[16] Sun, Qingyu, Xi Zhao, and Xinning Lin. "Design of a Hardware-Software Co-designed Real-Time Machine Learning System for Big Data Streams." Proceedings of the 2nd International Symposium on Integrated Circuit Design and Integrated Systems. 2025.