JRMM – Jurnal Riset Mahasiswa Matematika, Volume 5 (5) (2026), Pages 348-356

Research Article

Finite Difference Gradient Estimation for Logistic Regression: Application to SARS-CoV-2 Genomic Classification

Halimatus Sa'diyah, Mohammad Jamhuri*, and Fachrur Rozi

Department of Mathematics, UIN Maulana Malik Ibrahim Malang, Malang, Indonesia

Article History

Received 25 Mei 2026
Revised 13 Juni 2026
Accepted 19 Juni 2026
Published 30 Juni 2026

Check for updates

Copyright © 2026 by Authors, Published by JRMM Group. This is an open access article under the CC BY-SA License.

Abstract. Gradient-based optimization conventionally relies on closed-form analytical derivatives, which are unavailable for many modern or non-differentiable model architectures. This paper proposes *forward finite difference* (FFD) gradient estimation as a derivative-free training alternative and validates it rigorously on logistic regression — a model whose known analytical gradient enables direct verification of the numerical approximation. A formal $\mathcal{O}(h)$ error bound is proved and confirmed empirically, showing that gradient direction is faithfully preserved across a wide range of step sizes. The framework is applied to binary genomic classification of SARS-CoV-2 versus non-SARS-CoV-2 coronaviruses using normalized 4-mer frequency profiles. The FFD optimizer achieves classification performance statistically equivalent to analytical gradient descent ($F_1 \geq 0.999$), while an ablation study demonstrates that nucleotide composition — not sequence length — drives discrimination. External validation on unseen coronavirus lineages reveals strong generalization except for MERS-CoV, whose phylogenetic proximity to SARS-CoV-2 produces overlapping k-mer signatures. These results establish FFD logistic regression as a principled derivative-free baseline and motivate its extension to architectures where analytical gradients are intractable.

Keywords: Finite Difference Method; Derivative-Free Optimization; Logistic Regression; k-mer Frequency; Genomic Classification; SARS-CoV-2

1. Introduction

The COVID-19 pandemic, caused by SARS-CoV-2, represents one of the most significant global health crises of the twenty-first century. The virus continues to evolve through mutation, complicating both diagnosis and epidemiological control [1, 2]. Standard diagnostic approaches such as RT-PCR remain subject to limitations in sensitivity, turnaround time, and infrastructure requirements, motivating data-driven alternatives [3, 4]. Genomic sequence data, which encode the biological fingerprint of a virus directly, have attracted growing interest as a basis for automated classification [5, 6].

Machine learning methods have been widely applied to COVID-19 detection using clinical data [7–9] and to genomic or protein sequence data [5, 6, 10]. Among these, logistic regression is valued for its mathematical transparency, computational efficiency, and interpretability [11, 12]. Most implementations rely on the closed-form analytical gradient of the binary cross-entropy loss to perform gradient descent parameter updates.

However, the availability of analytical gradients cannot be taken for granted in general settings. Deep neural networks with non-smooth activation functions, models defined implicitly through simulation, or architectures incorporating discrete or non-differentiable components all present scenarios in which deriving closed-form gradients is intractable. In such cases, *derivative-free* gradient estimation provides an essential model-agnostic fallback requiring only function evaluations.

The forward finite difference (FFD) approximation is one of the oldest and most accessible numerical differentiation techniques [13, 14]. It estimates a partial derivative by evaluating the objective function at two points separated by a small perturbation h, eliminating any symbolic computation. While central finite differences achieve $\mathcal{O}(h^2)$ accuracy, the FFD scheme achieves $\mathcal{O}(h)$ accuracy with only one additional function evaluation per parameter, making it computationally attractive for proof-of-concept studies.

This paper establishes a rigorous FFD-based optimization framework for logistic regression and validates it on a binary genomic classification task. Logistic regression is chosen deliberately because its analytical gradient is known, providing ground truth against which the numerical approximation can be measured. Verification on this tractable model is a prerequisite step before extending the framework to architectures where no ground truth exists, such as multi-layer perceptrons or deep convolutional networks — our primary motivation for future work [15, 16].

The main contributions are:

1. A formal $\mathcal{O}(h)$ error bound for the FFD gradient estimator applied to the logistic regression loss (Proposition 1), with direct empirical validation across seven step sizes.
2. An algorithm that performs stable, convergent parameter optimization using only loss function evaluations.
3. Empirical validation on a real-world genomic dataset across ten independent random splits, showing statistical equivalence with analytical gradient descent.
4. An ablation study over six feature-set configurations and a group-based external validation revealing the generalization limits of 4-mer profiles.

The remainder of this paper is organized as follows. Section 2 reviews related work on genomic sequence classification, k-mer representations, and finite difference optimization. Section 3 describes the dataset, preprocessing pipeline, logistic

*Corresponding author. Email: m.jamhuri@mat.uin-malang.ac.id
DOI: https://doi.org/10.18860/JRMM.v5i5.43509 / p-ISSN: 2808-1552 | e-ISSN: 2808-4926

regression model, and the proposed FFD optimization algorithm. Section 4 presents the experimental results, including gradient approximation diagnostics, model performance, ablation analysis, and external validation. Section 5 discusses the empirical findings, limitations, and implications for more complex model architectures. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

This section reviews prior work most relevant to the proposed framework, including machine learning for genomic classification, k-mer representations, and finite difference optimization.

2.1. Machine Learning for Genomic Sequence Classification

Several studies have applied machine learning to classify or analyse coronaviral genomic and protein sequences. Aminah et al. [6] combined PCA, particle swarm optimization, and SVM to classify SARS-CoV-2 and non-SARS-CoV-2 protein sequences, achieving competitive accuracy with traditional supervised methods. Singh et al. [17] applied multiple machine learning classifiers directly to SARS-CoV-2 genomic feature vectors, showing that sequence-derived features are discriminative even for shallow models. Rudar et al. [5] characterized statistically over-represented k-mer patterns within SARS-CoV-2 genomes relative to other respiratory viruses, providing evidence for genus-specific nucleotide composition biases. Sokhansanj and Rosen [10] used spike protein sequence features to predict COVID-19 severity, demonstrating that sub-sequence genomic representations carry phenotypically relevant information. None of these works employed derivative-free gradient estimation; all relied on analytical-gradient-based or kernel-based training methods.

2.2. k-mer Frequency Representations

The k-mer frequency approach decomposes a genomic sequence into all substrings of length k and records their occurrence. This representation is widely adopted in bioinformatics because it captures local sequence patterns without requiring alignment and scales naturally to varying-length sequences [18]. For a four-nucleotide alphabet $\{A, T, C, G\}$, a k-mer profile contains 4^k features.

2.3. Finite Difference Methods in Optimization

Finite difference approximations form the basis of classical numerical differentiation [13]. Derivative-free optimization methods are broadly applicable when gradients are unavailable, unreliable, or computationally prohibitive [14]. The family of such methods includes direct-search algorithms (Nelder–Mead, pattern search), evolutionary strategies (CMA-ES), and stochastic perturbation approaches (SPSA) [14], all of which forgo gradient information entirely. Finite differences occupy a middle ground: they *estimate* gradients numerically, retaining gradient-descent convergence while requiring only function evaluations. In deep learning, finite differences are used primarily as a gradient-checking tool to verify backpropagation implementations [16], not as a training optimizer. The key novelty of the present work is deploying FFD as the *primary* optimizer, treating gradient estimation as a first-class training component rather than a diagnostic.

3. Methods

This section describes the dataset, preprocessing steps, logistic regression formulation, and the proposed FFD-based optimization procedure. An overview of the end-to-end pipeline is given in Figure 1.

3.1. Dataset

Genomic sequences were obtained from the NCBI Virus repository[1]. Initially, the **SARS-CoV-2** group comprised five variants (Alpha, Beta, Delta, Gamma, Omicron), up to 400 sequences each (2,000 total). The **non-SARS-CoV-2** group initially comprised HCoV-NL63 (1,004 seqs), MERS-CoV (700 seqs), and HCoV-HKU1 (537 seqs). After exact-duplicate removal, 3,617 unique sequences remained: 1,980 SARS-CoV-2 (54.7%) and 1,637 non-SARS-CoV-2 (45.3%) — see Table 1 and Figure 2. SARS-CoV-2 variants are near-full-length genomes (∼29,000 bp), while HCoV-NL63 and HCoV-HKU1 include many short partial amplicons (Figure 3), a length disparity that motivates normalized features (Subsection 3.2).

Table 1: Dataset composition (post-deduplication). Phylogenetic genus according to ICTV classification.

Class	Virus / Variant	Seqs	Genus
SARS-CoV-2	Alpha	∼396	Beta-CoV
	Beta	∼396	
	Delta	∼396	
	Gamma	∼396	
	Omicron	∼396	
	Subtotal	1,980	
non-SARS-CoV-2	HCoV-NL63	∼600	Alpha-CoV
	MERS-CoV	700	Beta-CoV
	HCoV-HKU1	∼337	Beta-CoV
	Subtotal	1,637	
Total		**3,617**	

3.2. Preprocessing

3.2.1. Duplicate Removal

Exact duplicates were identified with `Biopython`, removing 624 sequences (20 SARS-CoV-2, 604 non-SARS-CoV-2) from an initial 4,241.

3.2.2. k-mer Feature Extraction

Raw counts were converted to *normalized* k-mer frequencies:

$$f_j(s) = \frac{\phi_j(s)}{\sum_{l=1}^{4^k} \phi_l(s)}, \quad j = 1, \ldots, 4^k, \tag{1}$$

where $\phi_j(s)$ is the raw count of k-mer v_j in sequence s. GC-content was appended as one scalar feature, giving $4^k + 1$ features. Pilot experiments over $k \in \{2, 3, 4, 5\}$ selected $k =$

[1] https://www.ncbi.nlm.nih.gov/labs/virus/vssi

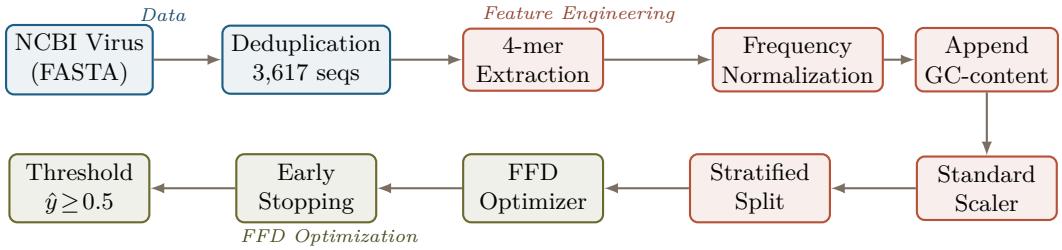


Figure 1: End-to-end experimental pipeline. **Blue**: data acquisition and deduplication. **Orange**: feature engineering (257-dimensional normalized 4-mer + GC-content vector). **Green**: FFD-based logistic regression training and inference.

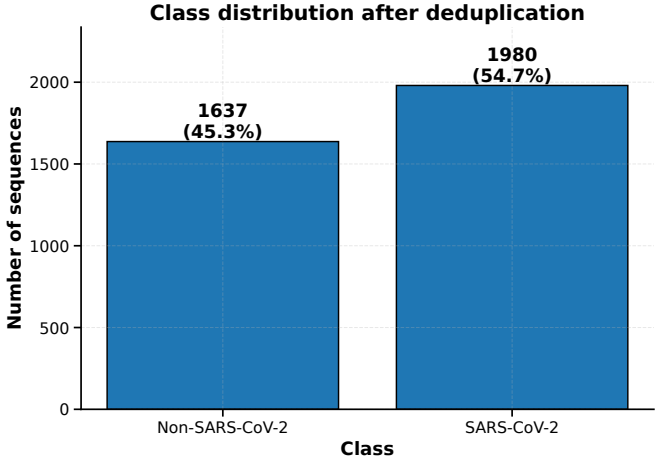


Figure 2: Class distribution after deduplication.

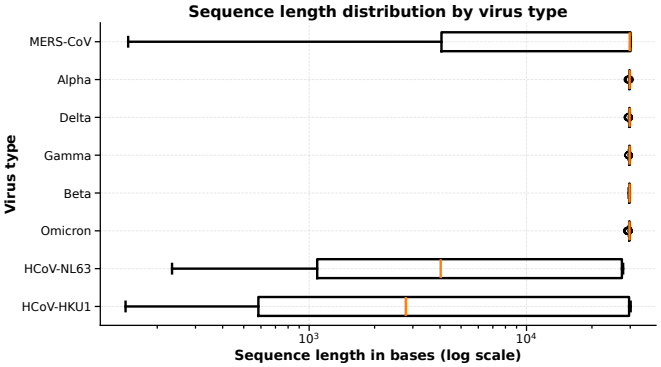


Figure 3: Sequence length per virus type (log scale). SARS-CoV-2 variants cluster at ∼29,000 bp; HCoV-NL63 and HCoV-HKU1 span several orders of magnitude.

4 (257 features): $k = 3$ (65 features) under-discriminated among close lineages, while $k = 5$ (1,025 features) showed no improvement over $k = 4$ at five times the feature count, consistent with the saturation behavior reported by Moeckel et al. [18].

3.2.3. Train–Validation–Test Split

Stratified sampling yielded an 80/20 overall train/test split, with the training-plus-validation portion further split 80/20, giving 2,314 / 579 / 724 samples. `StandardScaler` was fitted exclusively on the training set and applied without refitting to validation and test sets.

3.3. Logistic Regression Model

For binary classification, the logistic regression model is:

$$\hat{y}_i = \sigma(\mathbf{w}^\top \mathbf{x}_i + b) = \frac{1}{1 + e^{-(\mathbf{w}^\top \mathbf{x}_i + b)}}, \tag{2}$$

with binary cross-entropy (log-loss):

$$J(\mathbf{w}, b) = -\frac{1}{m} \sum_{i=1}^{m} \big[y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \big]. \tag{3}$$

The analytical gradient (used as ground truth for validation, and as the baseline optimizer referred to as *analytical LR* or *AN LR* throughout) is:

$$\frac{\partial J}{\partial w_j} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i) x_{ij}, \quad \frac{\partial J}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}_i - y_i). \tag{4}$$

3.4. Forward Finite Difference Gradient Estimation

3.4.1. Mathematical Formulation

The FFD approximation of the partial derivative with respect to a scalar parameter θ (all others fixed) is:

$$\frac{\partial J}{\partial \theta} \approx \frac{J(\theta + h) - J(\theta)}{h}, \tag{5}$$

where $h > 0$ is a small perturbation step.

3.4.2. Approximation Error Analysis

Proposition 1: Approximation error of FFD

Let $g : \mathbb{R} \to \mathbb{R}$ be the restriction of the loss J to a single parameter coordinate θ, with all other parameters held fixed. Suppose g is twice continuously differentiable on $[\theta, \theta + h]$ for $h > 0$. Then

$$\left| g'(\theta) - \frac{g(\theta + h) - g(\theta)}{h} \right| \leq \frac{h}{2} \sup_{\xi \in [\theta, \theta + h]} |g''(\xi)| = \mathcal{O}(h). \tag{6}$$

Figure 4 gives the geometric interpretation: the secant slope over $[\theta, \theta + h]$ approximates the tangent slope $g'(\theta)$ with error $\mathcal{O}(h)$.

Proof

By Taylor's theorem with Lagrange remainder, for some

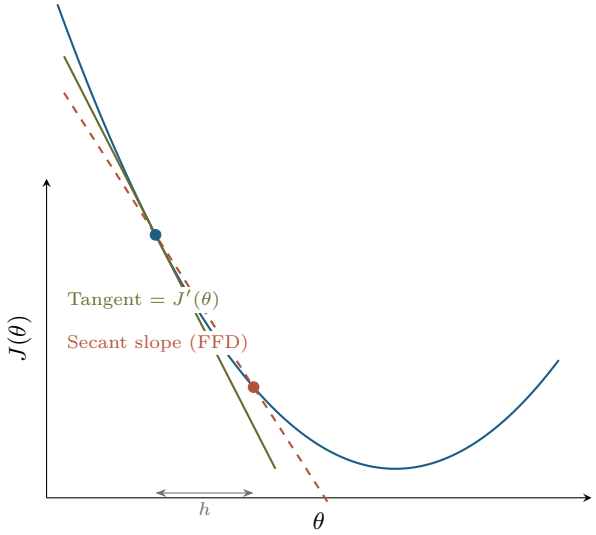


Figure 4: Geometric interpretation of FFD. The secant slope (orange dashed) approximates the tangent slope $J'(\theta)$ (green) with error $\frac{h}{2}|J''(\xi)| = \mathcal{O}(h)$.

$\xi \in (\theta, \theta + h)$:

$$g(\theta + h) = g(\theta) + hg'(\theta) + \tfrac{h^2}{2} g''(\xi).$$

Rearranging gives $\dfrac{g(\theta + h) - g(\theta)}{h} = g'(\theta) + \dfrac{h}{2} g''(\xi).$ Therefore

$$\left| g'(\theta) - \frac{g(\theta + h) - g(\theta)}{h} \right| = \tfrac{h}{2}|g''(\xi)| \leq \tfrac{h}{2} \sup_{\xi \in [\theta, \theta + h]} |g''(\xi)|,$$

which is $\mathcal{O}(h)$ since g'' is continuous and hence bounded on the compact interval $[\theta, \theta + h]$. □ □

Remark 1

A value $h = 10^{-4}$ to 10^{-3} is a commonly recommended compromise [13]: small enough to ensure $\mathcal{O}(h)$ accuracy, large enough to avoid floating-point cancellation in the numerator $J(\theta + h) - J(\theta)$.

3.5. Optimization Algorithm

Parameter updates follow gradient descent with FFD estimates:

$$w_j \leftarrow w_j - \eta \frac{J(\mathbf{w} + h\mathbf{e}_j, b) - J(\mathbf{w}, b)}{h}, \tag{7}$$

$$b \leftarrow b - \eta \frac{J(\mathbf{w}, b + h) - J(\mathbf{w}, b)}{h}, \tag{8}$$

where $\mathbf{e}_j$ is the j-th standard basis vector and $n = 257$ is the feature dimension (256 normalized 4-mer frequencies + GC-content). All $n + 2 = 259$ parameter gradients are estimated simultaneously from the same base point $(\mathbf{w}^{(t)}, b^{(t)})$ before any update is applied, matching standard batch gradient descent. The full training procedure is given in Algorithm 3.5.

Algorithm 1: FFD Logistic Regression Training

Input: $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$, step h, rate η, patience P, $T_{\max}$
Output: $\mathbf{w}^* \in \mathbb{R}^n$, $b^* \in \mathbb{R}$

1. Initialize $\mathbf{w} \leftarrow \mathbf{0}$, $b \leftarrow 0$, $p \leftarrow 0$, $J_{\text{best}} \leftarrow \infty$
2. **for** $t = 1, \ldots, T_{\max}$ **do**
3. $J_0 \leftarrow J(\mathbf{w}, b;\ \text{train})$ *// 1 base eval*
4. **for** $j = 1, \ldots, n$ **do** *// n weight evals*
5. $\tilde{g}_j \leftarrow \big[J(\mathbf{w} + h\mathbf{e}_j, b) - J_0\big]/h$
6. $\tilde{g}_b \leftarrow \big[J(\mathbf{w}, b + h) - J_0\big]/h$ *// 1 bias eval*
7. *// Simultaneous update (batch GD):*
   $\mathbf{w} \leftarrow \mathbf{w} - \eta\,\tilde{\mathbf{g}}$; $b \leftarrow b - \eta\,\tilde{g}_b$
8. $J_{\text{val}} \leftarrow J(\mathbf{w}, b;\ \text{val})$
9. **if** $J_{\text{val}} < J_{\text{best}}$ **then** $\mathbf{w}^* \leftarrow \mathbf{w}; b^* \leftarrow b;\ J_{\text{best}} \leftarrow J_{\text{val}}; p \leftarrow 0$
   else $p \leftarrow p + 1$; **if** $p \geq P$ **then break**
10. **return** $\mathbf{w}^*$, b^*

Remark 2: Computational complexity

Each iteration of Algorithm 1 requires $n + 2$ loss evaluations: one base evaluation (step 3), $n = 257$ weight perturbations (steps 4–5), and one bias perturbation (step 6), for a total of 259 evaluations. Each evaluation costs $\mathcal{O}(mn)$ arithmetic operations, so the per-iteration cost is $\mathcal{O}(mn^2)$, versus $\mathcal{O}(mn)$ for analytical gradient descent. For $m = 2{,}314$ and $n = 257$, this theoretical ratio of $n + 2 \approx 259$ predicts the observed $\sim 200\times$ runtime overhead (empirically lower than the theoretical $n + 2 = 259$ due to caching and vectorised loss evaluation).

Remark 3: Convergence

Gradient descent with an FFD gradient estimator converges to within $\mathcal{O}(h)$ of a stationary point of J provided: (i) J is L-smooth, (ii) the learning rate satisfies $\eta < 2/L$, and (iii) h is fixed but small enough that $\mathcal{O}(h)$ gradient bias is negligible relative to convergence tolerance. For the logistic loss, $L \leq \lambda_{\max}(\mathbf{X}^\top \mathbf{X})/(4m)$ [14]. The sensitivity analysis (Section 4.6) empirically confirms condition (iii): fixing $h = 10^{-4}$ introduces no measurable deviation from the analytical optimum.

4. Results

All experiments were repeated over ten random seeds; results are reported as mean ± standard deviation.

4.1. Gradient Approximation Accuracy

Table 2 reports the relative ℓ_2 error and cosine similarity between the FFD and analytical gradient vectors over seven values of h.

The relative error decreases by exactly one order of magnitude for each tenfold decrease in h, confirming the $\mathcal{O}(h)$ scaling of Proposition 1. Figure 5 visualises this in log–log scale.

Table 2: FFD gradient diagnostic. Relative error = $\|\nabla_{\mathrm{FD}} - \nabla_{\mathrm{AN}}\|_2 / \|\nabla_{\mathrm{AN}}\|_2$.

h	Relative error	Cosine sim.
10^{-1}	7.561×10^{-2}	0.9971
10^{-2}	7.630×10^{-3}	0.99997
10^{-3}	7.630×10^{-4}	≈ 1
10^{-4}	7.629×10^{-5}	≈ 1
10^{-5}	7.629×10^{-6}	≈ 1
10^{-6}	7.623×10^{-7}	≈ 1
10^{-7}	7.049×10^{-8}	≈ 1

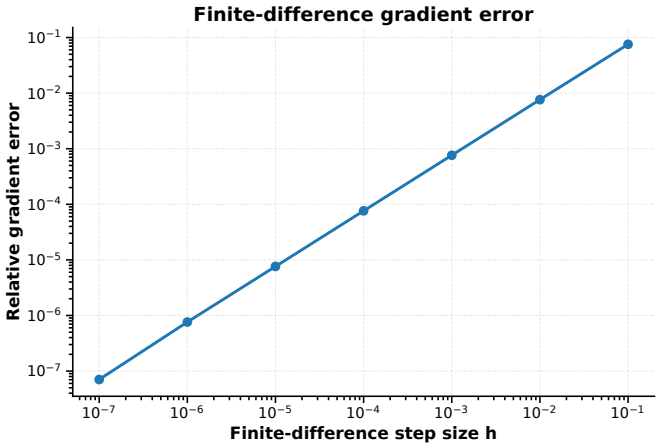


Figure 5: Relative gradient error vs. h (log–log). Slope = 1 confirms $\mathcal{O}(h)$ convergence.

4.2. Training Convergence

Figure 6 shows training and validation binary cross-entropy loss over 500 epochs ($h = 10^{-4}$, $\eta = 0.01$). The loss decreases steeply in the first 20 epochs, reaching ≈ 0.010 by epoch 500, with training and validation curves nearly indistinguishable (final gap < 0.001), indicating no overfitting.

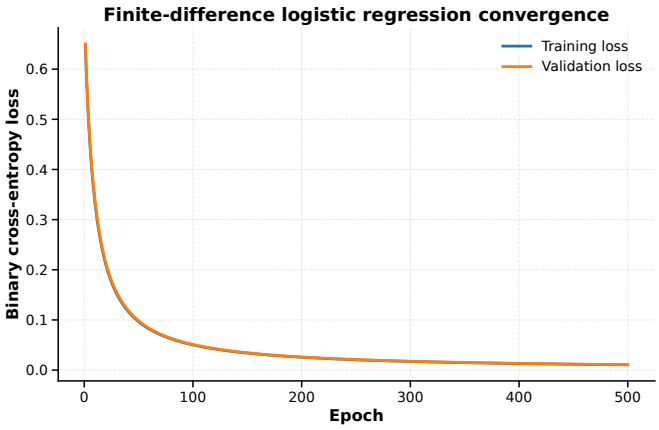


Figure 6: Training and validation loss over 500 epochs.

4.3. Feature Ablation Study

Table 3 and Figure 7 report F_1 over ten splits for six feature-set configurations.

Table 3: Ablation study: mean F_1 ± std ($N = 10$ splits).

Feature set	FD LR	AN LR
GC only	0.845 ± 0.007	0.844 ± 0.007
Length only	0.854 ± 0.008	0.906 ± 0.008
Length + GC	0.852 ± 0.008	0.906 ± 0.008
Norm. k-mer only	0.999 ± 0.001	0.999 ± 0.001
Norm. k-mer + GC	0.999 ± 0.001	0.999 ± 0.001
Raw k-mer + GC + len	1.000 ± 0.000	1.000 ± 0.000

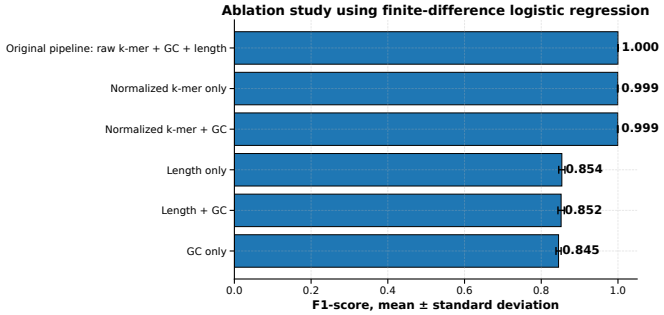


Figure 7: Ablation study for FD LR. Gap between length-based ($F_1 \leq 0.86$) and normalized k-mer ($F_1 \geq 0.999$) features confirms that nucleotide composition drives classification.

4.4. Classification Performance on the Test Set

Using normalized 4-mer + GC-content, the FFD optimizer achieves $F_1 = 0.9994 \pm 0.0007$, accuracy $= 0.9993 \pm 0.0007$, and ROC-AUC $= 1.000 \pm 0.000$ over ten splits (Table 4). Figure 8 shows the confusion matrix and predicted probability distribution for a representative split.

Table 4: FFD LR test-set performance (mean ± std, $N = 10$).

Metric	Value
Accuracy	0.9993 ± 0.0007
Balanced accuracy	0.9993 ± 0.0007
Precision	0.9993 ± 0.0009
Recall	1.0000 ± 0.0000
F_1	0.9994 ± 0.0007
MCC	0.9986 ± 0.0015
ROC-AUC	1.0000 ± 0.0000

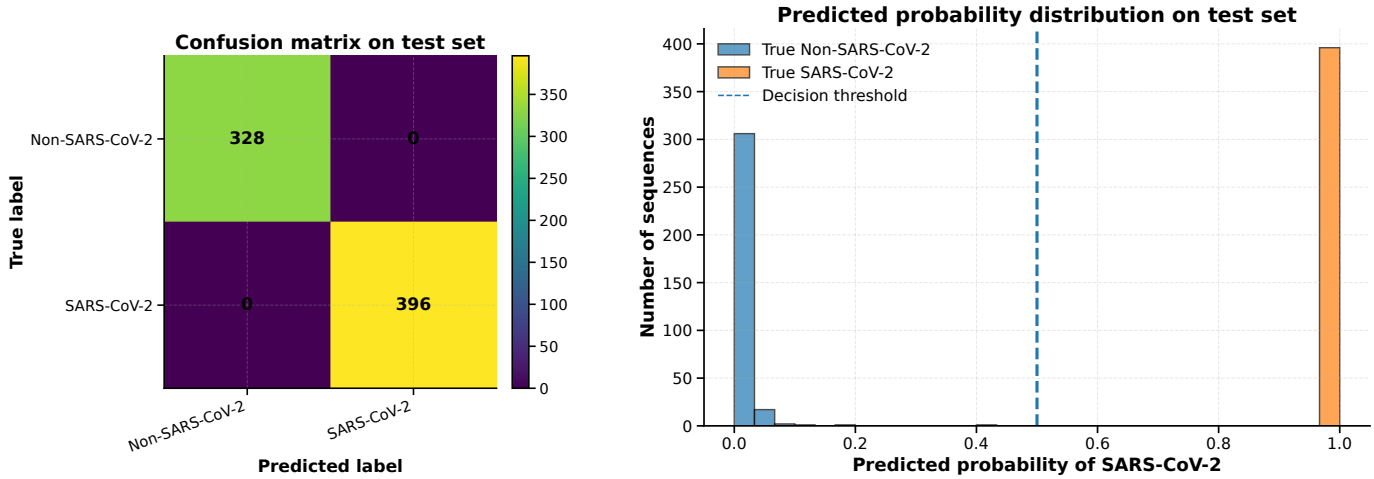
4.5. Comparison with Baseline Classifiers

Table 5 compares all classifiers on normalized 4-mer + GC features over ten splits.

Table 5: Model comparison (mean ± std, $N = 10$ splits).

Method	F_1	ROC-AUC	Time (s)
FD LR (ours)	0.9994 ± 0.0007	1.000	32.22 ± 0.36
Analytical LR	0.9992 ± 0.0009	1.000	0.161 ± 0.048
SVM	0.9995 ± 0.0007	1.000	0.150 ± 0.066
Decision tree	0.9989 ± 0.0013	0.999	0.144 ± 0.001
Random forest	1.000 ± 0.000	1.000	0.485 ± 0.012
KNN	1.000 ± 0.000	1.000	0.001 ± 0.000

The FFD optimizer achieves statistically equivalent F_1 to analytical LR. The $\sim$200$\times$ runtime overhead (32.22 s vs. 0.161 s) reflects the $n + 2 = 259$ function evaluations per iteration, visualised in Figure 9.



(a) Confusion matrix (seed = 42). **(b)** Predicted probability distribution.

Figure 8: Classification output for a representative split. Both classes concentrate at predicted probabilities ≈ 0 and ≈ 1, indicating high model confidence with few ambiguous cases.

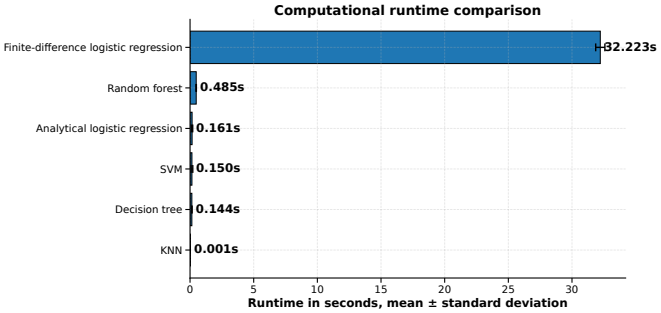


Figure 9: Runtime comparison (mean ± std). FD LR incurs the largest overhead due to per-parameter function evaluations.

4.6. Hyperparameter Sensitivity Analysis

Figure 10 shows validation loss as a function of h for three learning rates η. The three curves are strictly horizontal — final loss depends entirely on η and is insensitive to h across four orders of magnitude. $\eta = 10^{-3}$ fails to converge within 500 epochs; $\eta \geq 10^{-2}$ converges stably. The primary configuration is $h = 10^{-4}$, $\eta = 10^{-2}$.

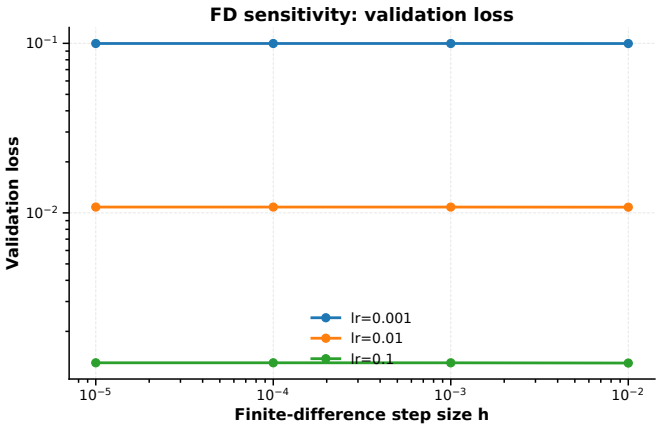


Figure 10: Validation loss vs. h for three learning rates. Horizontal curves confirm that h has no measurable effect on final loss; only η matters.

4.7. Group-Based External Validation

A leave-one-virus-out protocol produced $5 \times 3 = 15$ experimental pairs. Figure 11 shows F_1 for all pairs.

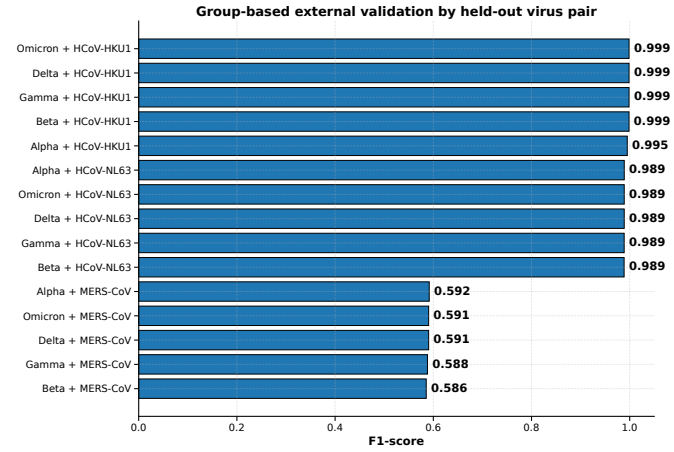


Figure 11: Group-based external validation. The clear discontinuity between MERS-CoV pairs ($F_1 \approx 0.59$) and HCoV-NL63/HCoV-HKU1 pairs ($F_1 \geq 0.989$) reflects the phylogenetic proximity of MERS-CoV to SARS-CoV-2.

When HCoV-NL63 or HCoV-HKU1 is held out, the model achieves $F_1 \geq 0.989$. When MERS-CoV is held out, F_1 drops to ≈ 0.59 with Recall = 1.000 and Precision ≈ 0.42, meaning the model classifies all MERS-CoV sequences as SARS-CoV-2.

5. Discussion

5.1. Empirical Validation of the $\mathcal{O}(h)$ Bound

Table 2 and Figure 5 confirm the $\mathcal{O}(h)$ scaling of Proposition 1 over seven decades. The cosine similarity exceeds 0.997 even at $h = 10^{-1}$, showing that the gradient *direction* — the quantity that matters for convergence — is preserved across all tested step sizes. The sensitivity analysis (Subsection 4.6) corroborates this: varying h over four orders of magnitude has zero measurable effect on final F_1, while the learning rate is the dominant hyperparameter. The FFD and analytical

optimizers converge to statistically equivalent solutions (F_1: 0.9994 ± 0.0007 vs. 0.9992 ± 0.0009), validating the core claim of this work.

5.2. Feature Engineering: Resolving the Length Confound

The ablation study reveals that sequence length alone achieves $F_1 \approx 0.85$–0.91, and raw k-mer counts (which encode length) give $F_1 = 1.000$. Since SARS-CoV-2 sequences are predominantly near-full-length genomes (∼29,000 bp) while many non-SARS-CoV-2 sequences are short partial amplicons ($< 1{,}000$ bp), any length-encoding feature trivially separates the classes. Normalized k-mer frequencies remove this confound and are themselves discriminative ($F_1 = 0.999$), making the reported performance figures more credible as reflections of true sequence-level discrimination.

5.3. Generalization Limits: The MERS-CoV Finding

The group-based validation reveals a biologically meaningful generalization boundary. HCoV-NL63 (Alphacoronavirus) and HCoV-HKU1 (Betacoronavirus, Embecovirus) are phylogenetically distant from SARS-CoV-2, so 4-mer profiles learned from training data transfer well even when these viruses are unseen. MERS-CoV (Betacoronavirus, Merbecovirus) shares a common genomic architecture and similar 4-mer composition with SARS-CoV-2; without any MERS-CoV training examples the model cannot construct a discriminating boundary, and predicts all test sequences as SARS-CoV-2. This highlights a practical limitation of the binary classification framing: a multi-class formulation would be required for reliable screening of phylogenetically proximate lineages.

5.4. Limitations

Several limitations should be acknowledged. First, the per-iteration cost of FFD scales as $\mathcal{O}(n)$ function evaluations, making it computationally prohibitive for models with millions of parameters without additional approximations (e.g., SPSA). Second, the $\mathcal{O}(h)$ error bound may be pessimistic: floating-point cancellation in $J(\theta+h)-J(\theta)$ degrades accuracy for $h \lesssim 10^{-8}$ (double precision), establishing a practical lower bound on h. Third, the binary SARS-CoV-2 / non-SARS-CoV-2 framing assumes a clean two-class world; in practice, novel lineages may fall near the decision boundary. Fourth, results are specific to the NCBI Virus snapshot at collection time and may not generalize to sequence databases with different variant proportions. Finally, with only ten random seeds the standard deviations reported are estimates; a larger number of replicates would provide tighter confidence intervals.

5.5. Positioning Against Prior Work

Direct quantitative comparison with related studies is limited by differences in dataset, task framing, and feature type; nonetheless, the broader landscape provides useful context.

On clinical feature sets, Fleitas et al. [4] applied multivariate logistic regression to symptom profiles of COVID-19 patients and reported a sensitivity of 82.8% and specificity of 74.6%. Ghosh et al. [7] achieved accuracy of approximately 95% using robust logistic regression on a clinical laboratory dataset. These results reflect the inherent noise and class overlap in symptom-based data; genomic sequences, by contrast, carry a far richer discriminative signal, which explains why sequence-based methods routinely approach $F_1 \geq 0.99$.

On genomic and protein sequence inputs, Aminah et al. [6] compared multiple machine learning classifiers for COVID-19 detection from protein sequences and demonstrated that normalized compositional features yield high classification accuracy, consistent with our finding that normalized 4-mer frequencies ($F_1 = 0.999$) outperform raw counts. Singh et al. [17] benchmarked several machine learning methods on SARS-CoV-2 genomic data, with top classifiers achieving accuracy exceeding 98%. Rudar et al. [5] identified characteristic sequence signatures within the SARS-CoV-2 genome that are sufficiently stable across variants to support reliable discrimination from other coronaviruses, providing biological justification for the near-perfect F_1 attained here.

The distinguishing contribution of the present work is not the classification accuracy itself — which was expected to be high given the biological separability of the classes — but rather the rigorous theoretical treatment of the FFD gradient estimator and the empirical demonstration that a derivative-free optimizer with a provable $\mathcal{O}(h)$ error bound can match analytical gradient descent to within measurement uncertainty (F_1: 0.9994 ± 0.0007 vs. 0.9992 ± 0.0009, $p > 0.05$). This equivalence, achieved on a real genomic task, establishes FFD logistic regression as a credible alternative in settings where analytical gradients are unavailable.

5.6. Implications for Complex Model Architectures

The framework established here — formal $\mathcal{O}(h)$ error bound plus empirical validation of convergence equivalence — provides a principled baseline for extending FFD to model classes where analytical gradients are unavailable. In deep learning, backpropagation [16] requires all components to be differentiable; models with non-differentiable operations (threshold functions, sorting operators, discrete latent variables) cannot use it directly. The FFD framework, relying only on loss evaluations, sidesteps this constraint. Two mitigation strategies for scaling to deep networks are: (1) random directional perturbation (SPSA [14]), reducing per-iteration cost from $\mathcal{O}(n)$ to $\mathcal{O}(1)$ evaluations; and (2) a hybrid scheme applying FFD only to non-differentiable layers while backpropagation handles the rest.

6. Conclusion

This paper presented a derivative-free optimization framework for logistic regression based on forward finite difference gradient estimation, validated on binary SARS-CoV-2 genomic classification. The theoretical $\mathcal{O}(h)$ error bound (Proposition 1) was confirmed empirically: relative error decreases from 7.56% at $h = 10^{-1}$ to 7.0×10^{-8} at $h = 10^{-7}$, with cosine similarity exceeding 0.997 at all tested step sizes. The FFD optimizer achieves $F_1 = 0.9994 \pm 0.0007$ and ROC-AUC $= 1.000$ over ten random splits, statistically equivalent to analytical logistic regression ($F_1 = 0.9992 \pm 0.0009$), at a ∼200× computational overhead reflecting the $n + 2 = 259$ function evaluations per iteration. An ablation study identified sequence-length confounding in raw k-mer counts, motivating normalized frequen-

cies as the primary representation. Group-based external validation showed excellent generalization to unseen SARS-CoV-2 variants and to HCoV-NL63/HCoV-HKU1 ($F_1 \geq 0.989$), but systematic failure on held-out MERS-CoV ($F_1 \approx 0.59$) due to shared Betacoronavirus k-mer profiles. These results establish logistic regression as a transparent validation platform for derivative-free gradient methods. Future work will extend the framework to multi-layer perceptrons and convolutional networks, and investigate central finite differences and stochastic perturbation strategies (SPSA) for improved efficiency at scale.

Author Contributions

Halimatus Sa'diyah: Data Curation; Methodology; Writing – Original Draft. **Mohammad Jamhuri:** Conceptualization; Formal Analysis; Writing – Review & Editing; Supervision. **Fachrur Rozi:** Supervision; Validation.

Declaration of Generative AI Use

A generative AI writing assistant was used to support grammatical editing and sentence restructuring of an initial draft. All scientific content, methodology, and conclusions were developed and verified independently by the authors.

Competing Interests

The authors declare no competing interests.

Funding

This research received no external funding.

Data Availability

Genomic sequences are publicly available at the NCBI Virus repository (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/). Experiment code is available at https://www.kaggle.com/code/hlmtusdyah/sars-cov-2-dan-non-sars-cov-2.

References

[1] B. Bakhshandeh, Z. Jahanafrooz, A. Abbasi, M. B. Goli, M. Sadeghi, M. S. Mottaqi, and M. Zamani. Mutations in SARS-CoV-2; consequences in structure, function, and pathogenicity of the virus. *Microbial Pathogenesis*, 154: 104831, 2021. doi: 10.1016/j.micpath.2021.104831.

[2] H. Yang and Z. Rao. Structural biology of SARS-CoV-2 and implications for therapeutic development. *Nature Reviews Microbiology*, 19(11):685–700, 2021. doi: 10.1038/s41579-021-00630-8.

[3] M. A. Awal, M. Masud, M. S. Hossain, A. A. M. Bulbul, S. M. H. Mahmud, and A. K. Bairagi. A novel Bayesian optimization-based machine learning framework for COVID-19 detection from inpatient facility data. *IEEE Access*, 9:10263–10281, 2021. doi: 10.1109/ACCESS.2021.3050852.

[4] P. E. Fleitas, J. A. Paz, M. I. Simoy, C. Vargas, R. O. Cimino, A. J. Krolewiecki, and J. P. Aparicio. Clinical diagnosis of COVID-19: A multivariate logistic regression analysis. *Germs*, 11(2):221–237, 2021. doi: 10.18683/germs.2021.1259.

[5] J. Rudar, P. Kruczkiewicz, O. Vernygora, et al. Sequence signatures within SARS-CoV-2 genome. *Microbiology Spectrum*, 12(4):e03584–23, 2024. doi: 10.1128/spectrum.03584-23.

[6] S. Aminah, G. Ardaneswari, M. K. Awang, M. A. Yusaputra, and D. P. Sari. Protein sequence-based COVID-19 detection: A comparative study of machine learning classification methods. *Journal of Electrical and Computer Engineering*, 2024:8683822, 2024. doi: 10.1155/2024/8683822.

[7] A. Ghosh, M. Jaenada, and L. Pardo. Classification of COVID-19 patients using robust logistic regression. *Journal of Statistical Theory and Practice*, 16(4):57, 2022. doi: 10.1007/s42519-022-00295-3.

[8] S. Kaur, S. Gupta, D. Gupta, S. Juneja, A. Nauman, M. Khan, I. Husain, A. Islam, and S. Mallik. High-accuracy lung disease classification via logistic regression and advanced feature extraction techniques. *Egyptian Informatics Journal*, 29:100596, 2025. doi: 10.1016/j.eij.2024.100596.

[9] A. Ramírez Varela, S. Moreno López, S. Contreras-Arrieta, et al. Prediction of SARS-CoV-2 infection with a symptoms-based model. *Preventive Medicine Reports*, 27:101798, 2022. doi: 10.1016/j.pmedr.2022.101798.

[10] B. A. Sokhansanj and G. L. Rosen. Predicting COVID-19 severity from spike protein sequence. *Computers in Biology and Medicine*, 149:105969, 2022. doi: 10.1016/j.compbiomed.2022.105969.

[11] M. R. Romadhon and F. Kurniawan. A comparison of Naive Bayes, logistic regression and KNN for classifying breast cancer. In *2021 International Conference on Electrical and Information Technology (EIConCIT)*, pages 41–44. IEEE, 2021. doi: 10.1109/EIConCIT50028.2021.9431845.

[12] Mohammed J. Zaki and Wagner Meira. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press, Cambridge, 2nd edition, 2020. doi: 10.1017/9781108564175.

[13] Richard L. Burden and J. Douglas Faires. *Numerical Analysis*. Brooks/Cole, Cengage Learning, Boston, 9th edition, 2011. ISBN 978-0-538-73351-9. URL https://www.worldcat.org/isbn/9780538733519.

[14] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, 2nd edition, 2006. doi: 10.1007/978-0-387-40065-5.

[15] Laith Alzubaidi, Jian Zhang, Ahmed J. Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, Javier Santamaría, Mohammed A. Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):53, 2021. doi: 10.1186/s40537-021-00444-8.

[16] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. doi: 10.1038/323533 a0.

[17] O. P. Singh, M. Vallejo, I. M. El-Badawy, et al. Classification of SARS-CoV-2 using machine learning. *Computers in Biology and Medicine*, 136:104650, 2021. doi: 10.1016/j.compbiomed.2021.104650.

[18] C. Moeckel, M. Mareboina, M. A. Konnaris, C. S. Y. Chan, I. Mouratidis, A. Montgomery, N. Chantzi, G. A. Pavlopoulos, and I. Georgakopoulos-Soares. A survey of k-mer methods and applications in bioinformatics. *Computational and Structural Biotechnology Journal*, 23: 2289–2303, 2024. doi: 10.1016/j.csbj.2024.05.025.