Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15691
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Efficient CNN-Based Classification of SARS-CoV-2 Spike Gene Sequences Using Alignment-Free Encoding

Rengga Anggarah[1)], Ernawati[2)] ,Widhia KZ Oktoeberza[3)]
[1)2) 3)] Informatics Study Program, Faculty of Engineering, University of Bengkulu, Bengkulu, Indonesia
[1)]anggarahrengga@gmail.com, [2)]ernawati@unib.ac.id, [3)] widhiakz@unib.ac.id

Submitted :Dec 12, 2025 | **Accepted** : Jan 02, 2026 | **Published** : Jan 04, 2026

Abstract: The COVID-19 pandemic caused by SARS-CoV-2 continues to challenge the global health system through the emergence of various variants with genetic characteristics that affect vaccine transmission and effectiveness. Conventional identification methods such as *Whole-Genome Sequencing* (WGS) have high accuracy but are constrained by significant cost and time. Most classification studies today still rely on complex hybrid architectures such as CNN-LSTM or image-based representations that increase computational load. This study aims to develop an *efficient and lightweight pure Convolutional Neural Network* model based on *alignment-free encoding* to classify five *Variant of Concern* (VOC) variants of SARS-CoV-2 (Alpha, Beta, Delta, Gamma, and Omicron) with an exclusive focus on the Spike gene sequence. The dataset consists of 5,000 Spike gene sequences that are represented using *integer encoding* and standardized with *zero-padding*. CNN *proposed Lightweight* architecture consists of four 1D convolution layers with a total of approximately 1.6 million parameters. The test results show that the model achieves excellent performance with an overall accuracy of 98.93%. The precision, *recall*, and *F1-score* values averaged 0.99, while the analysis of the ROC curve showed AUC values above 0.99 for all variants. This approach has proven to be efficient and effective, offering a fast, scalable, and resource-efficient solution to support real-time genomic surveillance systems in future pandemic mitigation.

Keywords: *SARS-CoV-2, Genomic Variants, Spike Gene, Convolutional Neural Network, Integer Encoding, Genomic Classification*

INTRODUCTION

Pandemic *COVID-19* caused by viruses *Severe Acute Respiratory Syndrome Coronavirus 2* (*SARS-CoV-2*) has caused a global health crisis since the end of 2019. This virus undergoes continuous mutations and produces various new variants with different transmission characteristics and clinical severity. The World Health Organization (WHO) classifies several variants as *Variants of Concern* (VOC), namely *Alpha*, *Beta*, *Gamma*, *Delta*and *Omicron*. Genetic changes in these variants, especially in genes *Spike*, has been shown to affect the ability of the virus to transmit and adapt to the immune system, so that the process of detecting and classifying variants is an important aspect in pandemic control.

Although Whole-Genome Sequencing (WGS) has high accuracy, this method is constrained by high cost, time, and computational complexity, making it difficult to apply massively in health emergencies.

Technological developments *Artificial Intelligence* (AI), in particular *Deep learning*, providing new opportunities to efficiently analyze large amounts of biological data. One of the architectures that is widely used in complex data analysis is *Convolutional Neural Network* (CNN). *CNN* able to extract important patterns and features automatically from raw data, so it doesn't require any processing *alignment* complex as in WGS.

Previous studies prove the effectiveness of CNN for the classification of the SARS-CoV-2 genome (Câmara et al., 2022) achieve >95% accuracy with an alignment-free approach, (Whata & Chimedza, 2021) reported 99.95% accuracy using CNN-BiLSTM, while (Awe et al., 2023) achieved >98% accuracy using the CNN-LSTM hybrid model on the Spike gene. Research conducted by (Awe et al., 2025) Developing a hybrid model ***CNN*-LSTM** to classify dominant variants *SARS-CoV-2* (*Omicron*, *Delta*, *Beta*, *Gamma*, and *Alpha*) based solely on protein gene sequences *Spike* virus. This model achieved very high accuracy (99.91% on test data), indicating that the gene-focused model *Spike* It is accurate enough for variant classification.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15691
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Previous studies predominantly employ hybrid CNN-LSTM architectures or visual encoding approaches (CGR, k-mer images), which increase computational overhead. This study fills the gap by demonstrating that a simplified integer-encoded CNN can achieve equivalent classification performance (>98% accuracy) while preserving calibration interpretability. The main contributions are: (1) pure CNN with alignment-free encoding; (2) probability calibration analysis; (3) visualization-based interpretability of learned genomic patterns.

This study aims to develop an efficient and minimalist alignment-free encoding-based Convolutional Neural Network (CNN) model to classify five Variants of Concern (VOCs) of SARS-CoV-2 in a balanced manner, with an exclusive focus on the Spike gene sequence as the main area of mutation, using a large dataset (5,000 sequences) from NCBI. Significantly different from previous studies that generally relied on hybrid models (such as CNN-LSTM), federated learning, or image-based representation, this study presents a novelty by optimizing a pure CNN architecture based on efficient alignment-free integer encoding for the classification of Spike gene sequences, simpler but proven to be able to achieve very high accuracy (98.93%), equivalent to or surpass more complex models. This end-to-end alignment-free approach without complicated feature engineering offers a fast, scalable, and computationally resource-efficient Spike gene sequence-based variant classification solution, making it an efficient alternative to conventional methods such as WGS to support real-time variant monitoring. In addition to performance demonstrations, the study also integrates probability calibration analysis and visualization*filter*to uncover biological interpretability, an aspect that has rarely been explored in previous CNN studies of the SARS-CoV-2 genome.

LITERATURE REVIEW

SARS-CoV-2 is a positive single-stranded RNA virus from the Coronaviridae family with a genome of about 29.9 kb that encodes the main structural proteins, namely Spike (S), Envelope (E), Membrane (M), and Nucleocapsid (N). Among these proteins, the Spike protein has a crucial role in the infection process because it interacts directly with the ACE2 receptor in human host cells. Mutations in the Spike gene are known to contribute to increased transmissibility, virulence, and the ability of viruses to evade immune responses, so this gene is the main focus in research on the classification of SARS-CoV-2 variants.

The World Health Organization (WHO) has designated several variants of SARS-CoV-2 as Variants of Concern (VOC), namely Alpha, Beta, Gamma, Delta, and Omicron. Each variant has distinctive mutation characteristics, such as N501Y in the Alpha, L452R and P681R variants in the Delta variant, as well as the accumulation of significant mutations in the Omicron variant that affect the rate of transmission and immune escape ability (Andre et al., 2023).

Conventional SARS-CoV-2 variant identification is generally carried out through *Whole-Genome Sequencing* (WGS) that has high accuracy, but requires significant time, cost, and computing resources. This condition encourages the development of computational approaches based on *deep learning* for the classification of variants of the viral genome. Method *deep learning* It has proven to be effective in handling large, complex biological data, including genomic sequences.

CNNs are effective for genomic classification due to their ability to automatically extract nucleotide patterns.

A number of previous studies have shown the effectiveness of CNN and architecture *deep learning* others in the classification of SARS-CoV-2 variants. These studies used data variations (complete genome or Spike genes), sequence representation techniques, and diverse model architectures, with relatively high accuracy performance. However, these differences in approach suggest that there is no consensus on the simplest, calibrated, and easily interpretable CNN architecture for the classification of viral genome variants. Table 1 summarizes the comparison of methodologies and results from the main studies in this field.

Table 1. Comparison of Previous Research Classification of SARS-CoV-2 Variants

Author (Year)	Data	Method	Sequence Representation	Accuracy
(Wang et al., 2022)	The complete genome	CNN	One-hot	~98%
(Ullah et al., 2022)	Gen Spike	TCN	Encoding sekuens	~88.36%
(Harikrishnan et al., 2022)	The complete genome	NeuROChaos	Chaos-based	Height
(de Souza et al., 2023)	Through viruses	CNN	Spatial representation	Increase
(Coutinho et al., 2023)	Through SARS-CoV-2	SAE vs CNN	Encoding numerik	CNN excels
(Azevedo et al., 2024)	The complete genome	CNN 1D	Augmentasi	>99%

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15691
e-ISSN : 2541-2019
p-ISSN : 2541-044X

(Chourasia et al., 2024)	Gen Spike	Federated CNN	Encoding sekuens	~93%
(Bezerra et al., 2024)	Through SARS-CoV-2	CNN	CGR (2D)	99.8%

Previous studies have confirmed the effectiveness of deep learning for the classification of the SARS-CoV-2 genome, but a simple CNN architecture with good calibration and interpretability is still relatively limited.

METHOD

This study implements a CNN-based systematic approach with alignment-free encoding for the classification of SARS-CoV-2 variants using Spike gene sequences. The research methodology framework includes several critical stages: data collection and pre-processing, model architecture design, training methodology, and performance evaluation. Research (Nguyen et al., 2016) popularized the early approach to classifying DNA sequences by treating them as text data and using *CNN* 1D. Use *One-Hot Vector* Here it is very important because it can retain the essential position information of each nucleotide, which is the basic principle for the input *CNN* in the *genomics.*

Spike gene prioritization is justified due to mutation concentration in receptor binding domain (RBD) and immune-escape regions, which are critical determinants of transmissibility and antigenic variation among SARS-CoV-2 variants (Andre et al., 2023).

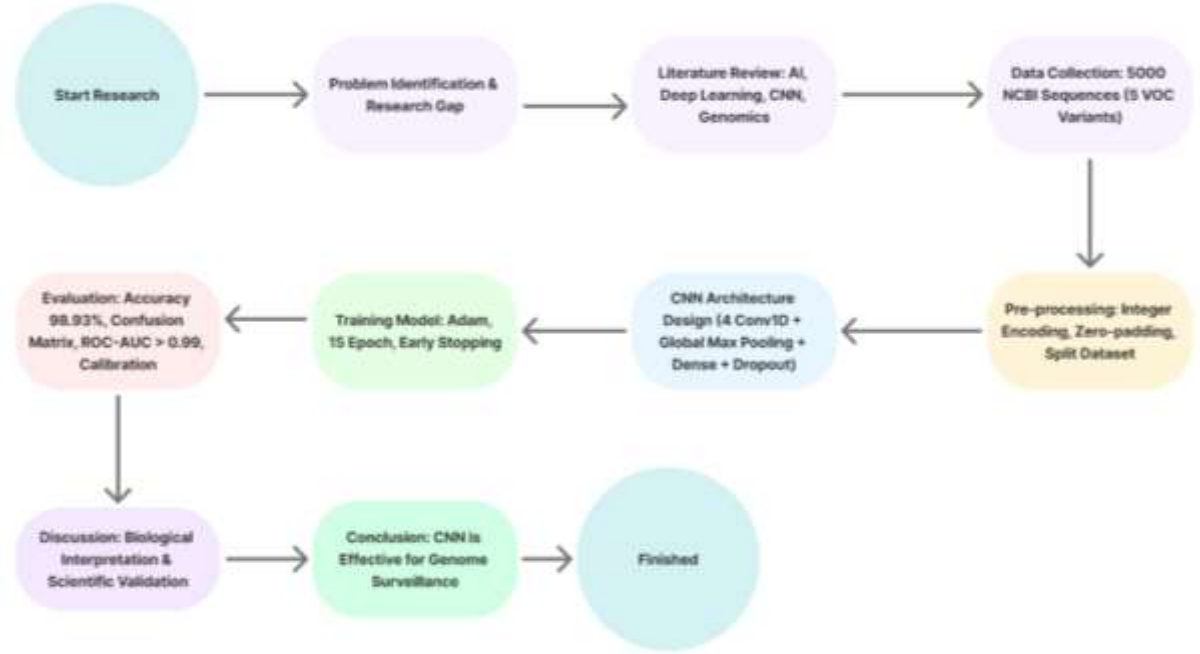


Figure 1. Research flow details

The research began with the identification of problems and the determination of *research gaps* related to the need for automated systems for *genome surveillance*. After that, a literature review was carried out on *AI, deep learning*, *CNN,* and *genomic analysis to strengthen the theoretical foundation. The researchers then collected 5000 genome sequences* from *NCBI* that included five VOC variants. The data is pre-processed through *integer encoding*, *zero-padding*, and dataset sharing. Furthermore, *the CNN architecture* was designed which consisted of four *Conv1D layers* followed by global max pooling, *dense layer* and *dropout*. The model was trained using *the Adam optimizer* for 15 *epochs* with *Early stopping* to ensure stable convergence. The training results were evaluated through accuracy, *Confusion Matrix*, *ROC-AUC* above 0.99, and model calibration. After evaluation, research proceeds to biological interpretation of the prediction and scientific validation. In the final stage, it was concluded that *CNN* was effectively used for *genome surveillance*, and the research was declared complete

CNN's selection was based on its ability to extract hierarchical features and translation-invariance advantages that have proven to be effective in a variety of domains (Li et al., 2022; Walz, 2023; Zhao et al., 2024). Thus, this study chose *CNN* due to its ability to automate hierarchical feature extraction from sequence data, translation-invariance advantages, and the maturity of deep learning technologies (*deep learning*) in a variety of domains. The architecture selection procedure includes model identification-*baseline* (e.g. *LeNet, AlexNet, VGG, ResNet*) that is

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15691
e-ISSN : 2541-2019
p-ISSN : 2541-044X

proven in previous studies, then the adaptation of the architecture to the context of the sequence *by SARS-CoV-2*. This method is justified by citing these sources and supporting the relevance of model selection in this study." In research (Nerkar & Kimbahune, 2024) discuss various techniques for coding DNA sequences (*Encoding Methods*) used for the model *machine learning* and *deep learning* like *CNN*Including *One-Hot Encoding* and *k-mer encoding*. Coding is an important step in converting sequence data into numerical inputs that can be understood by *CNN*.

The use of pure CNN was chosen because this architecture has been shown to be able to extract local patterns and nucleotide motifs that are important for the identification of mutations without requiring the long temporal structure that LSTM requires. CNN is also more stable, faster to train, and more computationally efficient for very long sequences such as viral genomes. This approach is reinforced by the use of *integer encoding*, which is intended to produce a minimalist and lightweight representation while maintaining the position of each nucleotide, so that the convolution process can study variant-specific mutation patterns without high complexity as in *one-hot encoding* or *K-MER*. According to (Potdar, 2017), Integer encoding (A=1, T=2, C=3, G=4) was chosen to produce a minimalist representation that was lightweight while maintaining the position of the nucleotide, allowing CNN to study mutation patterns without the complexity of one-hot encoding or k-mer. The selection of CNN in this study is also in line with the opinion (Rudin, 2022) This approach is in line with Rudin's (2022) principle that CNN effectively extracts features automatically, although it is often criticized as a black-box. According to (Sherstinsky, 2020), LSTMs are relatively insensitive to gap length so they are able to maintain long-term dependencies and recall values over varying time intervals, which makes them superior in processing more complex sequential information.

Data Collection and Pre-Processing

The dataset consists of 5,000 sequences *genome* which are evenly distributed across five variants *SARS-CoV-2* utama: *Alpha*, *Beta*, *Delta*, *Gamma*and *Omicron*, with each variant class consisting of 1,000 sequences, can be seen in Figure 2.

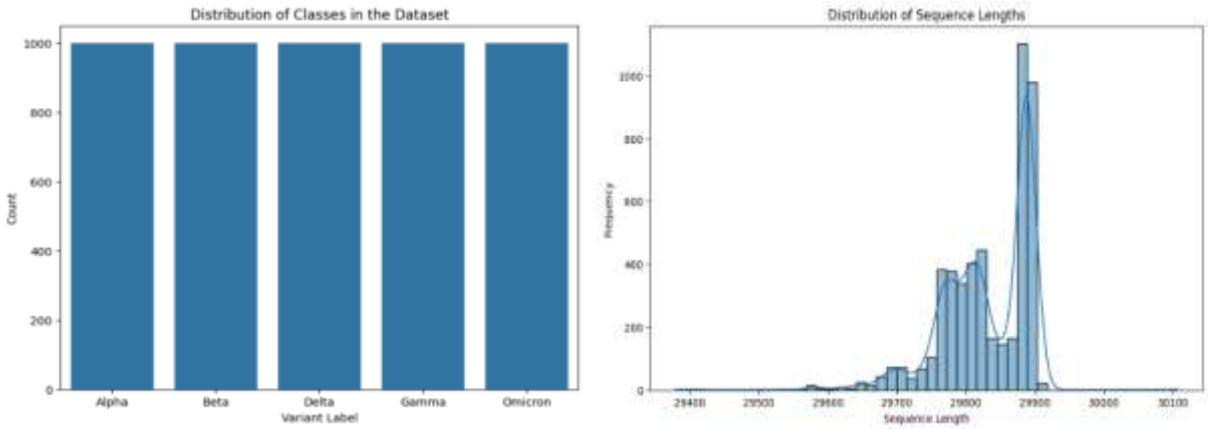


Figure 2. Dataset Class Distribution and Sequence Length Distribution

Sequence data were obtained from FASTA-format files downloaded from *NCBI* and undergoing a comprehensive pre-processing process. The nucleotide sequence is converted into numerical representations using a simple integer coding scheme where A=1, T=2, C=3, G=4, and other characters=0.

Integer coding was chosen to evaluate whether simplified coding retains discriminating power while reducing model complexity, as well as to consider computational feasibility for resource-constrained laboratories without relying on high-dimensional representations such as one-hot coding.

Given the considerable variation in the length of the sequence, with a maximum length of 30.107 nucleotides, it was carried out *zero-padding* to ensure uniformity of input dimensions across all sequences. The dataset is then strategically partitioned into training sets (70%, 3,500 sequences), validation (15%, 750 sequences), and *testing* (15%, 750 sequences) using *stratified sampling* to maintain a balanced distribution of classes across all partitions. Figure 2 also illustrates the length distribution of the sequence, highlighting the significant variations in length that require pre-processing *padding*. The maximum sequence length of 30,107 nucleotides underscores the computational challenges that the model architecture overcomes.

Model Architecture

Architecture *CNN* Designed specifically to handle sequence data *Genomics with configuration* presented in Table 2:

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15691
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Table 2 CNN lightweight architecture specifications

Layer (Type)	Output Shape	Estimate Configuration	Parameter
conv1d	(None, 30101, 64)	64 filters, kernel=7, activation=ReLU	512
max_pooling1d	(None, 10033, 64)	pool_size=3, stride=3	0
conv1d_1	(None, 10027, 128)	128 filters, kernel=7, activation=ReLU	57,472
max_pooling1d_1	(None, 3342, 128)	pool_size=3, stride=3	0
conv1d_2	(None, 3336, 256)	256 filters, kernel=7, activation=ReLU	229,632
max_pooling1d_2	(None, 1112, 256)	pool_size=3, stride=3	0
conv1d_3	(None, 1106, 512)	512 filters, kernel=7, activation=ReLU	918,016
max_pooling1d_3	(None, 368, 512)	pool_size=3, stride=3	0
global_max_pooling1d	(None, 512)	Global Max Pooling	0
dense	(None, 512)	512 units, activation=ReLU	262,656
dropout	(None, 512)	rate=0.5 (standard assumption)	0
dense_1	(None, 256)	256 units, activation=ReLU	131,328
dropout_1	(None, 256)	rate=0.5 (standard assumption)	0
dense_2	(None, 5)	5 units (Output), activation=Softmax	1,285

This model is a 1D Convolutional Neural Network (CNN) architecture designed for the classification of 5 categories of data. The model pipeline starts with four feature extraction blocks, where each block consists of a Conv1D layer with a 7-sized kernel and a ReLU activation whose number of filters gradually increases (64, 128, 256, to 512 filters). Each convolution process is followed by MaxPooling1D with *pool size* 3 to reduce the temporal dimension of the data.

Once feature extraction is complete, the model uses GlobalMaxPooling1D to summarize the most prominent features into a single vector of 512 size. The final stage is *Fully Connected Layer* consisting of two Dense layers (512 and 256 units) equipped with Dropout to prevent *overfitting*. This entire process is terminated by a Dense output layer with 5 units and Softmax activation to generate a classification probability, with a total of 1,600,901 parameters trained.

Training Configuration

The model is compiled using *optimizer Adam* with the *loss categorical cross-entropy* and accuracy as the main evaluation metric. The training was carried out for 15 *Epoch* with *batch size* 32. Usage 15 *Epoch* and *batch size* 32 was chosen because this value is commonly used in training *CNN* for data *genome* and is proven to provide a balance between gradient stability and convergence speed. Research *CNN through SARS-CoV-2* by (Gadelha et al., 2022) and (Azevedo et al., 2024) also showed that the model could achieve optimal convergence within ≤20 *Epoch* with *batch size* 32–128, up to configuration 15 *Epoch* and *batch size* 64 is consistent with previous research practices. *Early stopping callback* Implemented with Patience 5 *Epoch*Monitor *Validation loss* to prevent *overfitting* and restore the best model weight. The training process utilizes *Validation* Set for tuning *hyperparameter* and model selection.

Performance Metrics

A comprehensive evaluation was conducted using multiple metrics including accuracy, *loss*, *Confusion Matrix*, *classification report* (*precision*, *recall*, *F1-score*), *ROC curve*, and calibration analysis. Additional visualizations include visualizations *filter* from the layers *CNN* First, the analysis of the class distribution, the sequence length distribution, and the prediction probability distribution for each variant class.

RESULT

The proposed CNN model shows excellent performance in variant classification *SARS-CoV-2*, achieving extremely high accuracy and robust generalization across all evaluated metrics.

CNN Baseline Performance

Figure 3 shows a near-perfect classification with *true positive rate* very high in all variants: *Alpha* (149/150 = 99.3%), *Beta* (148/150 = 98,7%), *Delta* (148/150 = 98,7%), *Gamma* (149/150 = 99.3%), and *Omicron* (146/150 = 97.3%). Accuracy and recall for each variant are in the range of 97-99% (Table 3), with *F1-score* The average reaches 0.99. Minimal misclassification occurs mainly between *Delta* and *Omicron* (a total of 5 cases out of 300

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15691
e-ISSN : 2541-2019
p-ISSN : 2541-044X

samples of both variants), which have biological relevance due to the convergence of mutations in the receptor binding domain

Table 3 Performance Metrics Baseline Detail Classification

Varian	*Precision*	*Recall*	*F1-score*	Support
Alpha	0.99	0.99	0.99	150
Beta	0.98	0.99	0.99	150
Delta	0.97	0.99	0.98	150
Gamma	0.99	0.99	0.99	150
Omicron	0.99	0.97	0.98	150
Overall	**0.99**	**0.99**	**0.99**	**750**

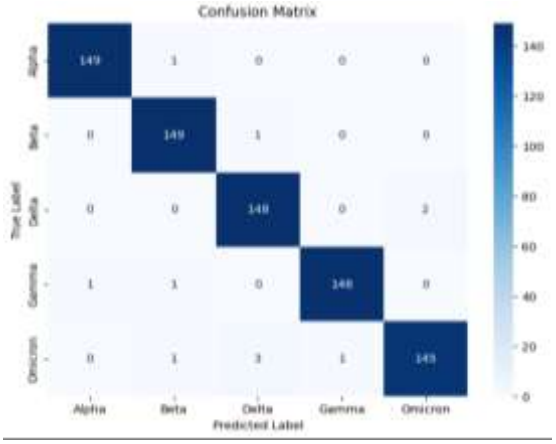


Figure 3 Confusion Matrix Baseline

In-depth analysis *off-diagonal elements* reveals biologically informative error patterns. The biggest mistake occurs between *Delta* and *Omicron* (*Delta→Omicron*: 2 cases; *Omicron→Delta*: 3 cases), which can be explained by a convergent substitution at a critical position: mutation *N501Y* who attended the *Omicron* and *L452R* Delta characteristics both affect interactions with *ACE2* receiver. At the nucleotide sequence level, this convergence results in overlapping k-mer patterns, causing ambiguity in *feature space* CNN learned. Other minor errors (*Alpha→Beta*: 1; *Gamma→Beta*: 1; *Omicron→Gamma*: 1) can be associated with *shared mutations* on *lineages* or noise in the sequencing data. Importantly, there are no variants that show *systematic misclassification pattern*, indicates a discriminative learning model *features genuine* for each class. *Absence of confusion* between certain pairs (e.g., no Alpha and Delta errors) suggest the genetic distance between these variants is large enough to allow linear separability within the decision boundaries of the softmax layer.

Performance of the CNN Lightweight Model

The CNN Lightweight model showed excellent classification performance with an overall accuracy of 98.93% on the test data. The average precision, recall, and F1-score values were in the range of 0.99, which indicates the model's ability to classify the five SARS-CoV-2 variants consistently and balanced. These results prove that the CNN architecture simplified with alignment-free integer encoding is still able to maintain high performance even when the complexity of the model is reduced.

Analisis *Confusion Matrix* CNN Lightweight

The CNN Lightweight model shows excellent classification performance with an overall accuracy of 98.93%. As seen in Table 4, this model produces an average precision and recall of 0.99 for all variants, with a consistent F1-score at the same number. The Alpha and Gamma variants showed perfect performance with precision, recall, and F1-score reaching 1.00, while the Delta variant had slightly lower precision at 0.97 but still maintained a high recall of 0.99. The Omicron variant shows the opposite pattern with a precision of 0.99 and a recall of 0.97, indicating a slight challenge in correctly identifying all Omicron samples. Overall, this lightweight model proves that a simpler architecture is still capable of achieving a very high level of accuracy in the classification of SARS-CoV-2 variants.

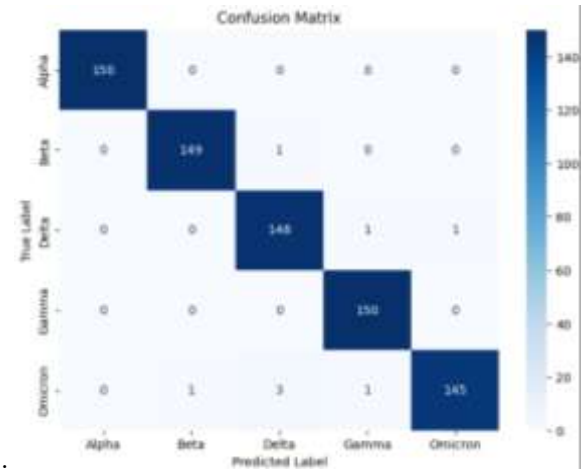
Table 4 Performance Metrics Detailed Classification CNN Lightweight

Varian	*Precision*	*Recall*	*F1-score*	Support
Alpha	1.00	1.00	1.00	150
Beta	0.99	0.99	0.99	150

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15691
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Delta	0.97	0.99	0.98	150
Gamma	0.99	1.00	0.99	150
Omicron	0.99	0.97	0.98	150
Overall	0.99	0.99	0.99	**750**

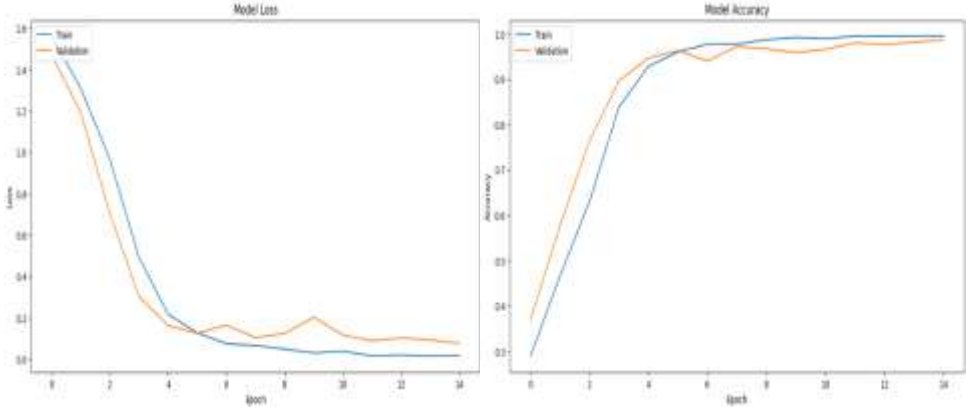


Gambar 4 *Confusion Matrix* CNN Lightweight

The Confusion Matrix in Figure 4 shows near-perfect classification capabilities with a highly accurate prediction distribution on the main diagonal. The Alpha variant achieved perfect classification with 150 of the 150 samples correctly identified with no errors at all. The Beta and Gamma variants also perform very well with only 1 misclassification each (149/150 correct). The Delta and Omicron variants had slightly more errors with 148 and 146 correct predictions out of 150 samples. An interesting pattern of error was seen in the Delta-Omicron relationship where there were 2 Delta samples that were misclassified as Omicron and 4 Omicron samples that were misclassified as Delta, confirming previous findings that these two variants have genetic similarities that can lead to ambiguity in classification. In total, only 11 errors from 750 testing samples showed an unusually high level of model accuracy.

CNN Lightweight Training Dynamics

The training curve in Figure 5 shows a stable and efficient convergence process over 15 epochs. The rapid convergence since the early epoch occurred because the SARS-CoV-2 variant has a very different mutation pattern in the Spike gene, so CNN can quickly recognize its genomic traits. This phenomenon is consistent with the findings (Azevedo et al., 2024) and (de Souza et al., 2023). The accuracy graph shows a rapid improvement in the early epochs, reaching training accuracy above 95% in the 5th epoch and steadily increasing until it is close to 99% in the later epoch. Validation accuracy follows a similar pattern without showing significant signs of overfitting, with minimal gaps between training and validation accuracy. The loss curve shows a consistent decline in training losses from an initial value of around 0.5 to close to 0.05 in the final epoch. Validation loss also decreased steadily although with slight minor fluctuations, which is an indication that the model is learning well without experiencing excessive overfitting. The stability of this curve confirms the effectiveness of the training configuration with the Adam optimizer, batch size 32, and the implemented early stopping mechanism.



*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15691
e-ISSN : 2541-2019
p-ISSN : 2541-044X

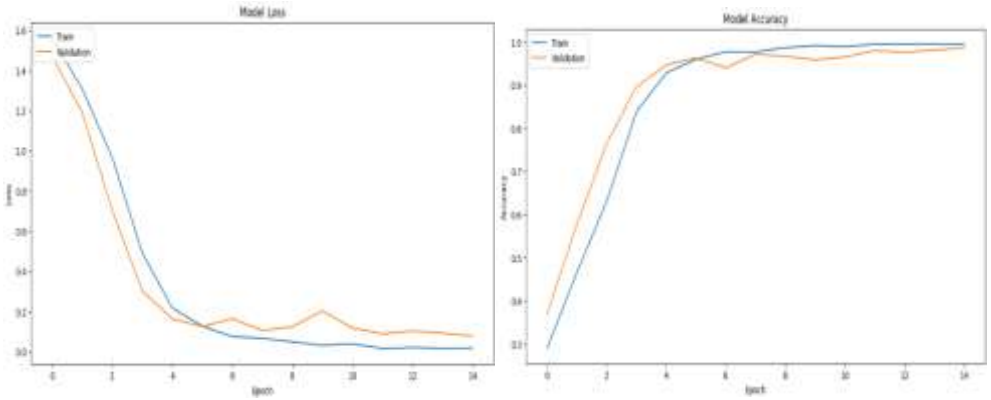


Figure 5 *Model Accuracy* and *CNN Lightweight* Loss Model

CNN Lightweight *ROC* Curve Analysis

The Receiver Operating Characteristic (ROC) curve in Figure 6 demonstrates the excellent model discriminating ability for all variants. Each variant shows an ROC curve approaching the upper left corner with an Area Under the Curve (AUC) value above 0.99, indicating that the model has an exceptional ability to distinguish each class from the others. The Alpha, Beta, Gamma, and Delta variants achieved an almost perfect AUC close to 1.00, suggesting there was practically no overlap in the distribution of prediction scores between positive and negative classes for these variants. The Omicron variant, although slightly lower, still maintains an AUC above 0.99. The macro-average and micro-average curves shown also show consistently high aggregate performance, validating that the model is unbiased towards a particular class and has balanced classification capabilities across variants.

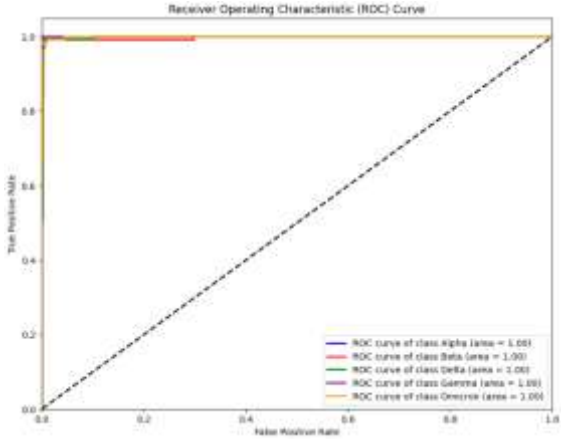


Figure 6 *ROC* Curve for All Classes CNN Lightweight Variant

CNN Lightweight Model Calibration

The calibration curve in Figure 7 shows an excellent match between the model's predictive probability and the actual frequency of positive events. The model's calibration line (blue line) follows very closely the perfect calibration line (diagonal), indicating that the probability predicted by the model is very well calibrated. When the model predicts a probability of 0.9 for a class, about 90% of that prediction is actually the class in question. The histogram at the bottom of the graph shows a concentrated prediction distribution at a high probability (close to 1.0), which is consistent with the model's high confidence level in the correct classification. This good calibration is critical for clinical applications because it ensures that the model's reported probabilities can be trusted and interpreted directly as a confidence level of the diagnosis, allowing for more informed medical decision-making.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15691
e-ISSN : 2541-2019
p-ISSN : 2541-044X

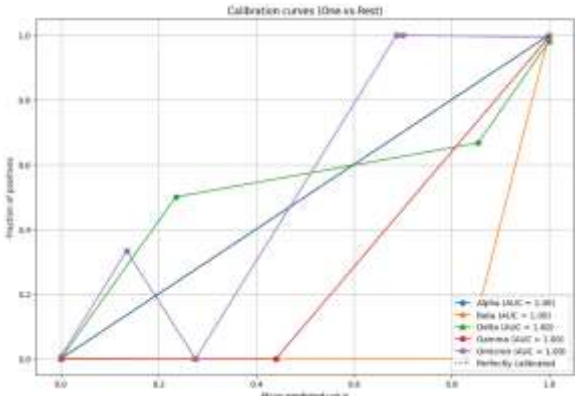


Figure 7 CNN Lightweight Calibration Curve

CNN Lightweight Architectural Insights

The filter visualization of the first convolutional layer in Figure 8 provides insight into the basic patterns learned by the model for feature extraction. These filters show variations in activation patterns that represent detectors for specific nucleotide motifs in the Spike gene sequence. Some filters display periodic patterns that may capture repeat sequences or codon patterns, while others show gradual patterns that can detect specific nucleotide transitions. The diversity of filter patterns indicates that the first layer successfully learns the representation of varied low-level features, which then combine in subsequent layers to form a more complex hierarchical representation. This visualization confirms that the model not only memorizes training data but actually learns biologically meaningful genomic structures, which contributes to the model's high generalization ability to test data.

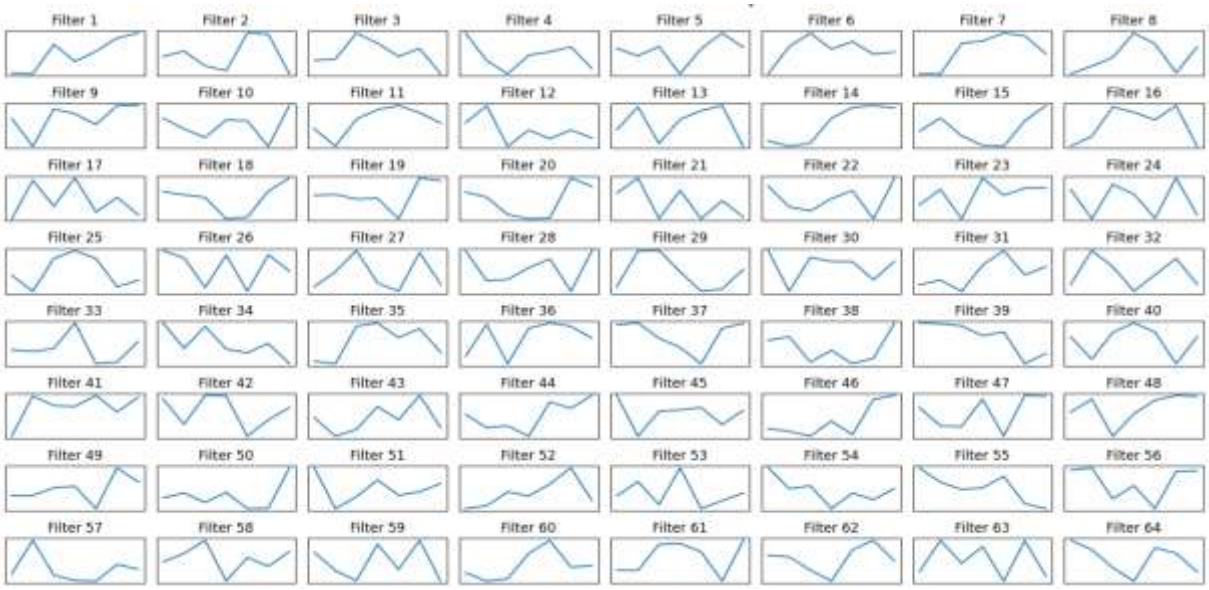


Figure 8 *Filters* from the *First* CNN Layer

CNN Baseline and CNN Lightweight Performance Comparison

Table 5. CNN Baseline and CNN Lightweight Performance Comparison

Model	Accuracy (%)	Precision (avg)	Recall (avg)	F1-score (avg)
CNN Baseline	98.53	0.99	0.99	0.99
CNN Lightweight	**98.93**	**0.99**	**0.99**	**0.99**

Table 5 presents a comprehensive comparison between the two model architectures developed in this study. CNN Lightweight achieved an accuracy of 98.93%, slightly higher than CNN Baseline which reached 98.53%. This 0.4% difference in accuracy suggests that both models have very comparable performance, with CNN Lightweight even outperforming it slightly despite having a simpler architecture. The average precision, recall and F1 score values for both models were identical at 0.99, indicating that the balance between accuracy and completeness of the predictions was well maintained in both architectures. These results reveal an important finding that increased model complexity is not always directly proportional to increased accuracy, and that more

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

efficient architectures can achieve performance equivalent to or even slightly better. This comparison supports the argument that for the classification of SARS-CoV-2 variants based on the Spike gene with integer encoding, the lighter CNN architecture is already quite optimal without the need for excessive complexity, offering advantages in terms of computational efficiency and inference speed while maintaining very high classification accuracy.

DISCUSSION

The results of the evaluation showed that the alignment-free encoding approach using integer encoding combined with the CNN architecture was able to effectively recognize SARS-CoV-2 variants. This experiment reveals an interesting performance dynamic between the CNN Baseline and CNN Lightweight models, reflecting the trade-off between architectural complexity and computational efficiency in capturing the genetic features of SARS-CoV-2 variants.

The CNN Baseline model developed in the early stages of the study showed excellent performance with an accuracy of 98.53% and a consistent precision, recall, and F1-score value of 0.99. The architecture with a progressive filter configuration (64-128-256) and four convolutional layers attests to its ability to extract complex hierarchical feature representations. Confusion Matrix analysis revealed that a major misclassification occurred between the Delta and Omicron variants (a total of 5 cases out of 300 samples), which can be biologically explained by the convergence of mutations in the receptor binding domain, specifically the N501Y mutation in Omicron and L452R in Delta which both affect interactions with ACE2 receptors. The stability of baseline model training is also well maintained despite the complexity reaching 36.7 million parameters, thanks to the implementation of a dropout rate of 0.5 and Global Max Pooling which functions as a natural regularization.

The development of CNN Lightweight with a simplified architecture yielded surprising findings. This model achieves an accuracy of 98.93%, slightly higher than the CNN Baseline, with identical precision, recall, and F1-score values of 0.99. The Alpha and Gamma variants showed perfect classification without errors, while the Delta-Omicron error pattern remained consistent with the baseline model. This performance improvement indicates that a more efficient architecture is able to reduce the risk of overfitting and improve generalization. The ROC curve with an AUC above 0.99 for all variants and an almost perfect calibration curve confirm that the lightweight model is not only accurate but also well calibrated, allowing for a reliable interpretation of predictive probabilities for clinical applications.

Comparatively, the performance of CNN Lightweight in this study offers a significant advantage over the hybrid models often found in the literature. These models achieve equivalent or even higher accuracy without requiring the architectural complexity of the recursive layer (LSTM/GRU) or image-based representation (CGR, k-mer images) that increase the computational load. This carries important practical implications: CNN Lightweight offers an optimal solution of inference speed, resource efficiency, and classification accuracy for real-time genomic surveillance systems, especially in computationally constrained laboratories. Filter visualization of the first convolutional layer revealed that the model was able to study biologically meaningful genomic patterns, including nucleotide motifs and sequence transitions specific to each variant.

These findings confirm that for the classification of SARS-CoV-2 variants based on the Spike gene sequence by integer encoding, the increase in model complexity is not always directly proportional to the increase in accuracy. Simpler and more efficient architectures can actually result in superior performance by reducing overfitting and improving generalization. Nonetheless, future research could focus on testing the robustness of models against sequence data that are partial, noisy, or derived from different sequencing technologies to simulate more challenging and realistic laboratory conditions.

LIMITATIONS OF THE RESEARCH

This research still has a number of limitations. First, the dataset only includes five publicly available VOC variants in the *NCBI*, so the model's performance on new or rare variants has not been verified. Second, the representation of integer encoding, while simple and efficient, has the potential to lose nucleotide context information that can be captured by encoding-based *k-mer* or *CGR*. Third, the model size is relatively large (≈36.7 million parameters), so implementation on low-resource devices requires advanced optimization. In addition, this study only used Spike sequence data; The model may require adaptation when used for the full genome or other viruses with different mutation structures.

CONCLUSION

This study successfully proved the effectiveness of an alignment-free encoding-based CNN architecture for the classification of SARS-CoV-2 variants using Spike gene sequences. Surprisingly, the CNN Lightweight model with a simpler architecture showed superior performance with an accuracy of 98.93%, surpassing the CNN Baseline of 98.53%. Both models maintained a consistent precision, recall, and F1-score values of 0.99 for all variants, demonstrating the optimal balance between prediction accuracy and completeness.

*name of corresponding author

CNN Lightweight proves that an efficient architecture with a smaller number of parameters is capable of achieving near-perfect classification, with the Alpha and Gamma variants achieving 100% accuracy with no errors at all. Minimal error patterns mainly occurred between the Delta and Omicron variants (4 cases out of 300 samples), which can be explained biologically through the convergence of mutations in the receptor binding domain. The Confusion Matrix analysis confirmed the absence of a systematic misclassification pattern, indicating that the model studied the discriminative features genuine for each variant.

The excellent calibration of the model, demonstrated by the near-perfect calibration curve and the ROC curve with an AUC above 0.99 for all variants, proves that the probability of prediction has a strong correlation with the actual outcome. This makes the model not only accurate but also trustworthy for clinical applications, where the interpretation of the diagnosis's confidence level is critical for medical decision-making. Filter visualization of the convolutional layer revealed that the model successfully studied biologically meaningful genomic patterns, including the specific nucleotide motifs that are the signature characteristics of each variant.

The main contribution of this study is the demonstration that a pure CNN approach with simple integer encoding can achieve very high classification performance (>98%) without the need for complex hybrid architecture (CNN-LSTM), image-based representation, or complex feature engineering commonly used in previous studies. The CNN Lightweight model offers a scalable, computationally resource-efficient, and fast solution for real-time genomic surveillance, making it an efficient alternative to conventional methods such as Whole-Genome Sequencing especially in computationally constrained laboratories. This study provides strong evidence that simplicity and efficiency in deep learning architecture design can produce superior performance while maintaining biological interpretability, supporting the implementation of more accessible variant detection systems for future pandemic mitigation.

Overall, this alignment-free encoding based CNN model on the Spike gene sequence has great potential to efficiently accelerate the genomic surveillance process of SARS-CoV-2, enable faster and more accurate variant identification at low computational costs, and support a more effective public health response in the face of pandemics.

REFERENCES

Andre, M., Lau, L. S., Pokharel, M. D., Ramelow, J., Owens, F., Souchak, J., Akkaoui, J., Ales, E., Brown, H., Shil, R., Nazaire, V., Manevski, M., Paul, N. P., Esteban-Lopez, M., Ceyhan, Y., & El-Hage, N. (2023). From Alpha to Omicron: How Different Variants of Concern of the SARS-Coronavirus-2 Impacted the World. *Biology*, *12*(9). https://doi.org/10.3390/biology12091267

Awe, O. I., obura, hesborn omwandho, Mwanga, M. J., & Evans, M. (2023). Enhanced Deep Convolutional Neural Network for SARS-CoV-2 Variants Classification. *BioRxiv*, 2023–2028.

Awe, O. I., Obura, H., Ssemuyiga, C., Mudibo, E., & Mwanga, M. J. (2025). *Enhanced deep Convolutional Neural Network for SARS-CoV-2 variants classification*. *September*, 1–16. https://doi.org/10.3389/frai.2025.1512003

Azevedo, K. S., de Souza, L. C., Coutinho, M. G. F., de M. Barbosa, R., & Fernandes, M. A. C. (2024). Deepvirusclassifier: a deep learning tool for classifying SARS-CoV-2 based on viral subtypes within the coronaviridae family. *BMC Bioinformatics*, *25*(1), 1–21. https://doi.org/10.1186/s12859-024-05754-1

Bezerra, G., Câmara, M., Prof, O., Augusto, M., & Fernandes, C. (2024). *Advanced Convolutional Neural Network Techniques for Classification of SARS-CoV-2 Variants and Other Viruses : A Study Using k -mers and Chaos Game Representation*.

Câmara, G. B. M., Coutinho, M. G. F., Silva, L. M. D. d., Gadelha, W. V. d. N., Torquato, M. F., Barbosa, R. de M., & Fernandes, M. A. C. (2022). Convolutional Neural Network Applied to SARS-CoV-2 Sequence Classification. *Sensors*, *22*(15), 1–15. https://doi.org/10.3390/s22155730

Chourasia, P., Murad, T., Tayebi, Z., Ali, S., Khan, I. U., & Patterson, M. (2024). Efficient Classification of SARS-CoV-2 Spike Sequences Using Federated Learning. *Communications in Computer and Information Science*, *2142 CCIS*, 80–96. https://doi.org/10.1007/978-3-031-63616-5_6

Coutinho, M. G. F., Câmara, G. B. M., Barbosa, R. de M., & Fernandes, M. A. C. (2023). SARS-CoV-2 virus classification based on stacked sparse autoencoder. *Computational and Structural Biotechnology Journal*, *21*, 284–298. https://doi.org/10.1016/j.csbj.2022.12.007

de Souza, L. C., Azevedo, K. S., de Souza, J. G., Barbosa, R. de M., & Fernandes, M. A. C. (2023). New proposal of viral genome representation applied in the classification of SARS-CoV-2 with deep learning. *BMC Bioinformatics*, *24*(1), 1–19. https://doi.org/10.1186/s12859-023-05188-1

Gadelha, W. V. N., Torquato, M. F., & Barbosa, R. D. M. (2022). *Sequence Classification*. 1–15.

Harikrishnan, N. B., Pranay, S. Y., & Nagaraj, N. (2022). Classification of SARS-CoV-2 viral genome sequences using Neurochaos Learning. *Medical and Biological Engineering and Computing*, *60*(8), 2245–2255. https://doi.org/10.1007/s11517-022-02591-3

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis,

*name of corresponding author

Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/TNNLS.2021.3084827

Nerkar, V., & Kimbahune, V. (2024). *Deep Learning Approaches in Genomic Analysis : A Review of DNA Sequence Classification Techniques*. *10*(2), 439–445.

Nguyen, N. G., Tran, V. A., Ngo, D. L., & Phan, D. (2016). *DNA Sequence Classification by Convolutional Neural Network*. *April*, 280–286.

Potdar, K. (2017). *A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers*. *175*(4), 7–9.

Rudin, C. (2022). *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x.Stop

Sherstinsky, A. (2020). *Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network*. *404*(March), 1–43.

Ullah, W., Ullah, A., Malik, K. M., Saudagar, A. K. J., Khan, M. B., Hasanat, M. H. A., AlTameem, A., & AlKhathami, M. (2022). Multi-Stage Temporal Convolution Network for COVID-19 Variant Classification. *Diagnostics*, *12*(11), 1–12. https://doi.org/10.3390/diagnostics12112736

Walz, W. (2023). *Machine Learning for Brain Disorders Series Editor*.

Wang, H., Tsinda, E. K., Dunn, A. J., Chikweto, F., Ahmed, N., Pelosi, E., & Zemkoho, A. B. (2022). *Deep learning forward and reverse primer design to detect SARS-CoV-2 emerging variants*. http://arxiv.org/abs/2209.13591

Whata, A., & Chimedza, C. (2021). Deep Learning for SARS COV-2 Genome Sequences. *IEEE Access*, *9*, 59597–59611. https://doi.org/10.1109/ACCESS.2021.3073728

Zhao, X., Wang, L., Zhang, Y., Han, X., Deveci, M., & Parmar, M. (2024). A review of convolutional neural networks in computer vision. In *Artificial Intelligence Review* (Vol. 57, Issue 4). Springer Netherlands. https://doi.org/10.1007/s10462-024-10721-6

*name of corresponding author