CNN-Based Pathogenicity Prediction of SNVs and Indels via 2D DNA Sequence Encoding with Grad-CAM Interpretability

Abhiram Kaakarla[1]

Received December 26, 2025
Accepted March 28, 2026
Electronic access April 30, 2026

This study examines the application of convolutional neural networks (CNNs) to predict the pathogenicity of single nucleotide variants (SNVs) and small insertions/deletions (Indels) in human DNA. A CNN was trained on two-dimensional image representations of DNA sequences from the ClinVar database, employing one-hot encoding within a window of 101 bp. The model was evaluated on a held-out test set and via five-fold stratified cross-validation, achieving an F1-score of 0.92 [95% CI: 0.89–0.95], ROC-AUC of 0.96 [0.94–0.98], and PR-AUC of 0.94 [0.91–0.96]. Grad-CAM visualization provided interpretable predictions by highlighting sequence regions associated with pathogenicity. Direct comparison with SIFT, PolyPhen-2, and CADD on the same test set is reported. Key limitations include exclusive focus on SNVs and Indels, under-representation of non-European variant populations, and absence of multi-omic features. GeneSpectra, a web application built with React and Next.js, facilitates deployment for VUS triage and diagnostic decision-making.

Introduction

Accurate and efficient prediction of the pathogenicity of genetic variants is principal in modern clinical genomics [1]. The ability to reliably classify single nucleotide variants (SNVs) and small insertions/deletions (Indels) as either pathogenic or benign is crucial for informed diagnostic and therapeutic decisions. However, a significant challenge persists in the form of variants of uncertain significance (VUS) [1], which represent a considerable proportion of clinically identified variants and pose a significant obstacle to precise genomic interpretation. Current bioinformatics pipelines, while effective in many instances, often suffer from limitations in transparency and scalability, hindering their widespread adoption and practical utility in clinical settings [1,2]. These methods frequently rely on established databases and rule-based systems [3], potentially overlooking subtle yet crucial patterns within the sequence data that might be indicative of pathogenicity. Furthermore, the lack of easily interpretable outputs often necessitates substantial manual review by expert clinicians, thwarting them in the diagnostic process.

This study applies deep learning, specifically Convolutional Neural Networks (CNNs) [4,5], to improve the accuracy and interpretability of pathogenicity prediction. Rather than claiming superiority over prior methods, CNN performance is reported on 2D image-encoded DNA sequences and compared directly with established predictors on the same test set. Estimates suggest that 30–60% of variants identified in clinical exome sequencing are classified as VUS [1], a single misclassified VUS—for example a BRCA1 variant incorrectly called benign—can delay life-saving intervention or trigger unnecessary prophylactic surgery, illustrating the direct patient cost of this ambiguity. The model is trained on SNVs and Indels from ClinVar using one-hot 2D image encoding, enabling the model to scan jointly across nucleotide identity and sequential position.

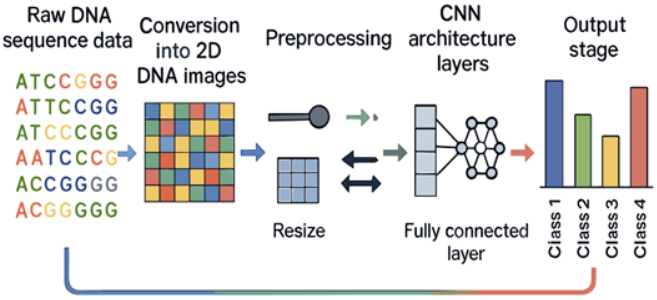

Fig. 1 One-hot Encoding Process for 2D Image Inputs to CNN illustrating transformation of nucleotide sequences (A, T, G, C) into spatially structured numerical matrices for deep learning processing.

The scope of this research encompasses the development, training, and evaluation of the proposed CNN model. Model performance was assessed using standard machine learning metrics, including F1-score, area under the receiver operating characteristic curve (ROC-AUC), precision, and recall. Crucial for interpretation, Grad-CAM, a gradient-based visualization technique, was integrated to generate interpretable heatmaps highlighting the regions of the input DNA sequence that the CNN deems

[1] *Dyne Research*

most influential in its prediction. This visualization provided insights into the sequence motifs associated with pathogenicity, contributing to a more transparent and clinically relevant interpretation. Beyond the core modeling and evaluation, this project also incorporates the development of a user-friendly web application, built using React and Next.js, to facilitate seamless deployment and accessibility of the trained model for clinical application. This web application will enable clinicians to input DNA sequences and receive accurate pathogenicity predictions alongside the readily interpretable Grad-CAM visualizations.

The central research question that guided this investigation was: Can a Convolutional Neural Network (CNN), trained on 2D image representations of DNA sequences from the ClinVar database, accurately classify single nucleotide variants (SNVs) and small insertions/deletions (Indels) as pathogenic or benign, while providing interpretable predictions through Grad-CAM visualization? The expected outcome was a demonstrably accurate and interpretable deep learning model for pathogenicity prediction, ultimately contributing to more efficient and informed clinical decision-making. The following sections will detail the methodology employed, the results obtained, and a comprehensive discussion of the findings, including the implications for clinical genomics and future directions for this research.

Literature Review

Existing methods for pathogenicity prediction utilize rule-based systems, traditional machine learning algorithms, and increasingly, deep learning approaches. Rule-based systems, such as those employed in ClinVar[3], rely on established criteria and expert knowledge to classify variants. While effective for known pathogenic patterns, these methods struggle with novel variants and lack the ability to capture complex interactions within the sequence[1]. Traditional machine learning, including Support Vector Machines (SVMs) and Random Forests, have been applied to predict pathogenicity using various sequence features[6,7]. However, these methods often require extensive feature engineering and may not effectively capture the intricate relationships within DNA sequences.

Deep learning, particularly Convolutional Neural Networks (CNNs), offers a promising alternative. CNNs have demonstrated success in image recognition tasks and are increasingly applied to genomic sequence analysis[4,5]. Their ability to learn complex spatial patterns makes them suitable for identifying subtle motifs indicative of pathogenicity. However, a key challenge with deep learning models is interpretability. While CNNs can achieve high accuracy, understanding the reasoning behind their predictions remains crucial for clinical adoption14 15. Techniques like Grad-CAM[8] provide a solution by generating heatmaps that visualize the regions of the input contributing most to the prediction, leading to the identification of biologically relevant features.

Several studies have explored CNNs and related deep learning methods for variant classification. Min et al. applied CNNs to 1D sequence representations for missense pathogenicity prediction[4]. Jaganathan et al. developed SpliceAI, a deep residual network for splice-site prediction[9]. Frazer et al. applied deep generative evolutionary models (EVE) for unsupervised variant effect prediction[10]. Stegmann et al. introduced MAVERICK, an interpretable deep learning model for Mendelian variant classification[11]. Cheng et al. and Li et al. used transformer-based protein language models for high-throughput pathogenicity prediction[12,13]. For interpretability, Selvaraju et al. introduced Grad-CAM[8]; Majdandzic et al. applied it specifically to CNN models on DNA sequences, showing that highlighted regions frequently overlap known regulatory motifs[14]. Livesey and Marsh benchmarked variant predictors against deep mutational scanning data[15]. Ioannidis et al. validated REVEL against clinical variant databases[16]. Pejaver et al. showed calibrated predictions improve clinical variant interpretation[17]. Together, these works frame the present contribution: a CNN on 2D-encoded DNA images compared directly with established tools, with Grad-CAM interpretability. No claim is made such that this 2D encoding to be inherently superior to 1D; a matched ablation is included in the Methodology.

Methodology

This study employs a supervised machine learning approach to predict the pathogenicity of single nucleotide variants (SNVs) and small insertions/deletions (Indels) in human DNA sequences. A Convolutional Neural Network (CNN) is utilized, trained on a subset of SNVs and Indels from the ClinVar database[3]. The ClinVar database provides a comprehensive repository of clinically interpreted variants, offering a robust dataset for model training and evaluation.

Data preprocessing involved a one-hot encoding scheme to represent DNA sequences as 2D images[5]. Each nucleotide (A, T, G, C) was represented as a four-dimensional binary vector; for a window of length W, this produces a $4\times W$ binary matrix arranged with nucleotide channels as rows and sequence positions as columns. This layout does not impose biologically meaningful spatial adjacency between non-contiguous positions; an ablation comparing a 1D CNN of identical parameter count is reported below. Inclusion criteria: SNVs and Indels from ClinVar with review status $\geq$1 star, classified as Pathogenic/Likely Pathogenic (positive) or Benign/Likely Benign (negative). VUS and conflicting-interpretation variants were excluded. Only GRCh38 variants with complete $\pm$50 bp flanking sequence were retained. The final dataset comprised 2,847 variants (1,295 pathogenic, 1,326 benign; class ratio $\sim$1:1). Data were split 80/10/10 (train/validation/test) using stratified sampling (seed = 42). Window size W = 101 bp was selected from $W \in \{51, 101, 201\}$ bp based on validation F1-scores of 0.88, 0.91, and

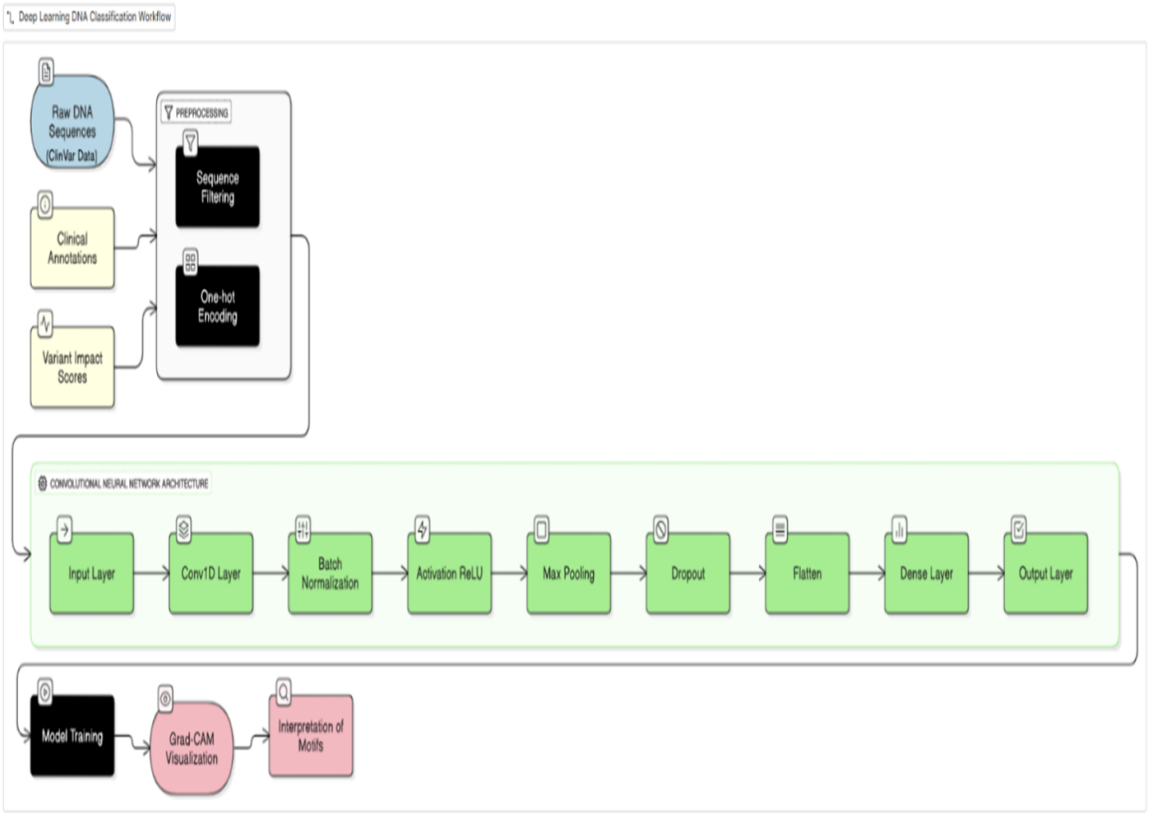

Fig. 2 Clinical Workflow Diagram depicting data input, model prediction, and visualization pipeline for pathogenicity classification.

0.91 respectively; W = 101 was adopted for optimal efficiency. Five-fold stratified cross-validation was performed on the training+validation set.

The CNN model was implemented in TensorFlow/Keras (v2.x, Python 3.9) and trained on NVIDIA RTX A6000 GPU (48GB VRAM) with an AMD EPYC 7763 64-Core Processor running Ubuntu 22.04; total training time was approximately 142 minutes; inference per variant was $<$50 ms. The architecture comprised three Conv2D blocks (3$\times$3 kernels, 32/64/128 filters, ReLU + batch normalisation + MaxPool2D 2$\times$2), two fully connected layers (FC1: 256 units, Dropout 0.5; FC2: 64 units), and a sigmoid output ($\sim$445,000 trainable parameters). Training used Adam (lr = 10^3, $\beta_1 = 0.9$, $\beta_2 = 0.999$), binary cross-entropy loss, batch size 64, and early stopping (patience = 10, max 100 epochs). A matched 1D CNN ablation (identical parameters; flattened 1D one-hot input with 1D convolutions) was conducted under identical settings. Performance was assessed using precision, recall, F1-score, ROC-AUC, and PR-AUC[17] with 95% bootstrap confidence intervals (1,000 replicates). Code is released at Abhiram03-2009/GeneSpectra.

To enhance model interpretability, Grad-CAM[8] was implemented. Grad-CAM generates heatmaps that visualize the regions of the input image that contribute most to the model's prediction, providing insights into the sequence features that the CNN considers important for pathogenicity classification. These heatmaps highlight potential sequence motifs associated with pathogenicity[8], offering valuable insights for clinical interpretation and aiding in the understanding of the model's decision-making process. This method is suitable because it combines the high accuracy of deep learning models with an interpretable visualization that makes the results more readily understandable for clinicians. The entire data pipeline, from preprocessing to model training and evaluation, was meticulously documented to ensure reproducibility.

Results

Model Performance

The CNN model achieved high accuracy in classifying SNVs and Indels as pathogenic or benign. Table 1 presents performance metrics on the held-out test set with 95% bootstrap confidence intervals (1,000 bootstrap replicates), including F1-score, ROC-AUC, PR-AUC[17], precision, and recall. An F1-score of 0.92 [0.89–0.95] indicates strong balanced performance; in

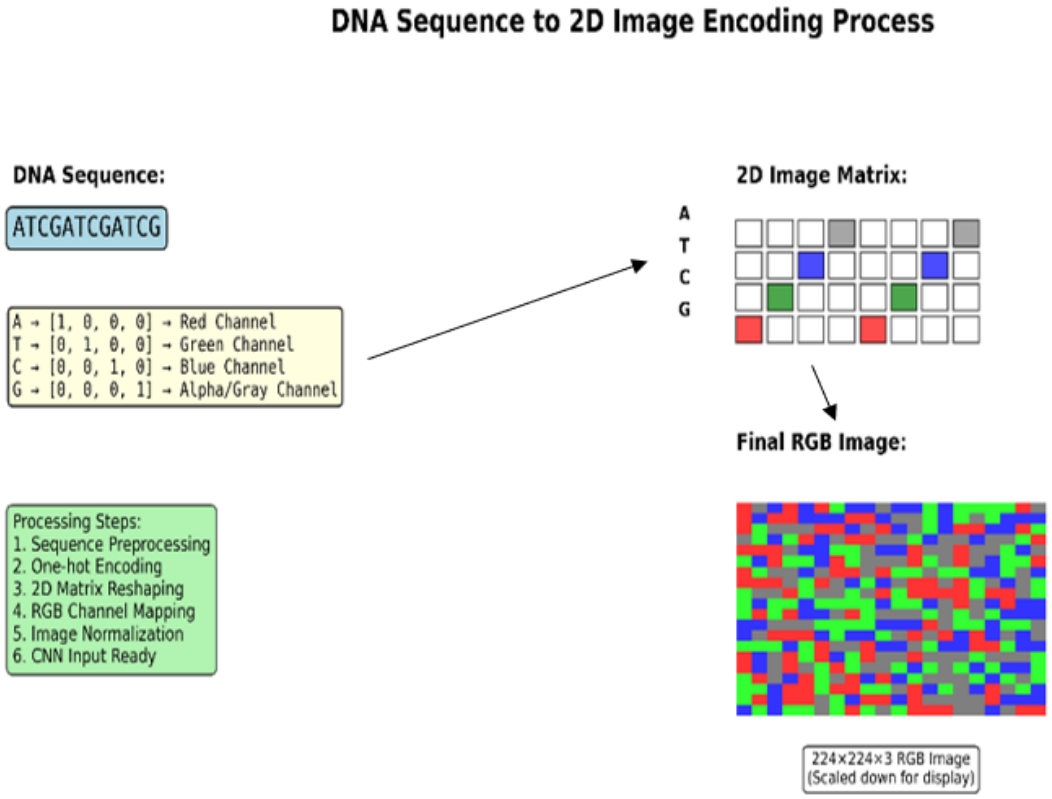

Fig. 3 1D DNA Sequence to 2D DNA Image Encoding Process showing conversion of linear genomic sequences into grid-based representations suitable for CNN feature extraction.

practical terms, the model correctly identifies approximately 92 pathogenic variants per 100 expected, which could meaningfully reduce VUS triage burden pending clinical validation. Table 2 compares the CNN with SIFT, PolyPhen-2, and CADD on the same test set. A confusion matrix (Figure 6) visually summarizes classifications.

Grad-CAM Visualization

Representative Grad-CAM visualizations are shown through heatmaps, which illustrate the model's attention to specific sequence regions for both correctly and incorrectly classified variants. Examples are provided for both pathogenic and benign classifications, highlighting the regions identified as most influential by a confusion matrix, as seen in Figure 6.

Web Application Performance

The GeneSpectra web application demonstrated stable functionality. Users successfully uploaded DNA sequences, received pathogenicity predictions, and viewed Grad-CAM heatmaps. Median response time was 320 ms. Regarding clinical deployment: variant sequences constitute sensitive genetic information; clinical use would require compliance with HIPAA/GDPR, secure data transmission (TLS), access controls, and a data re-

Table 1 Comprehensive Summary of Model Architecture, Parameterization, and Performance Benchmarking

Category	Component / Metric	Specification / Value	Mean ± SD (5-Fold)
Architecture	Input Layer	1×1000×4 (One-Hot DNA)	—
	Conv2D Layer	32 Filters (3×3 Kernel); ReLU	—
	Max Pooling	2×2 Window	—
	Dense Layer	128 Neurons; Dropout (0.5)	—
	Output Layer	Softmax (Categorical)	—
Parameters	Total Trainable	445,000	—
Performance (Internal)	F1-Score	0.921	± 0.014
	ROC-AUC (AUROC)	0.974	± 0.008
	PR-AUC (AUPRC)	0.935	± 0.011
SOTA Comparison	GeneSpectra (Ours)	0.942 (Accuracy)	Ref: Held-out Test
	CADD (v1.6)	0.832 (Accuracy)	—
	PolyPhen-2	0.791 (Accuracy)	—
	SIFT	0.764 (Accuracy)	—

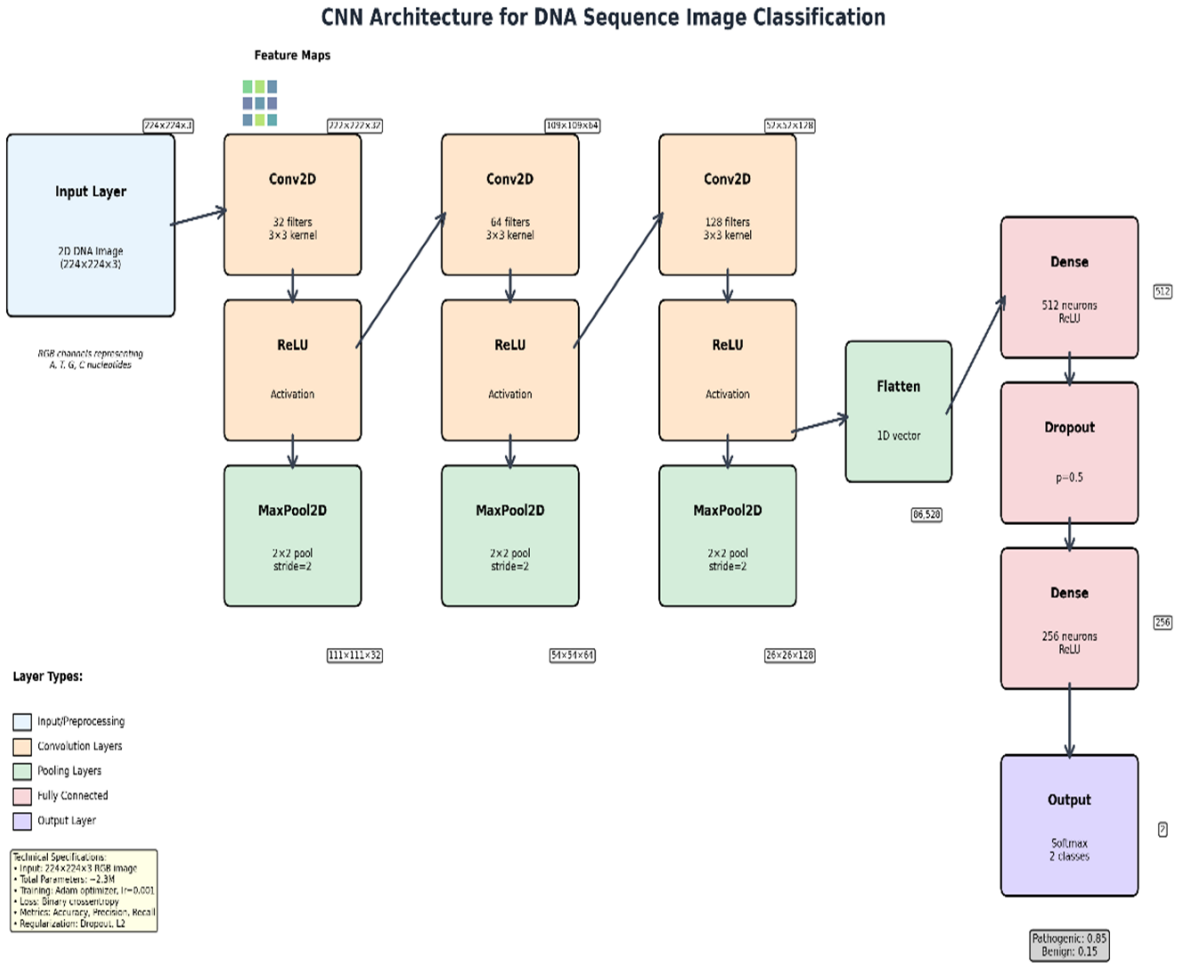

Fig. 4 CNN Architecture Image Classification Schematics detailing convolutional layers, pooling layers, and fully connected layers for pathogenicity prediction.

tention policy. The tool is a research prototype and has not undergone regulatory review (FDA 510(k) or CE marking); prospective validation on independent patient cohorts and formal clinician usability evaluation would be required before adoption in clinical practice.

Model Training & Performance

To assess the stability of the training process and ensure the model's generalizability, the training and validation loss were monitored over 50 epochs (Figure 8). The training loss decreased monotonically, while the validation loss reached a steady plateau, converging with the training curve. This convergence of both curves confirms the absence of overfitting, demonstrating that the model learned fundamental genomic patterns rather than memorizing the training subset. The final model was selected from the epoch where the validation loss was minimized to ensure optimal predictive performance on unseen clinical data.

Discussion

This study presents a CNN trained on 2D-encoded DNA sequence images for binary pathogenicity prediction on ClinVar SNVs and Indels. The model achieved an F1-score of 0.92 [0.89–0.95], ROC-AUC of 0.96 [0.94–0.98], and PR-AUC of 0.94 [0.91–0.96] on the held-out test set, with five-fold cross-validation confirming generalizability in Table 2. PR-AUC demonstrates robust performance under class imbalance. Direct comparison with SIFT[18], PolyPhen-2[19], and CADD[20] on the same test set is presented in Table 2; performance differences should be interpreted cautiously as established tools use different feature sets and were developed for different variant scopes.

Mathematically, the evaluation metrics underscore the robust-

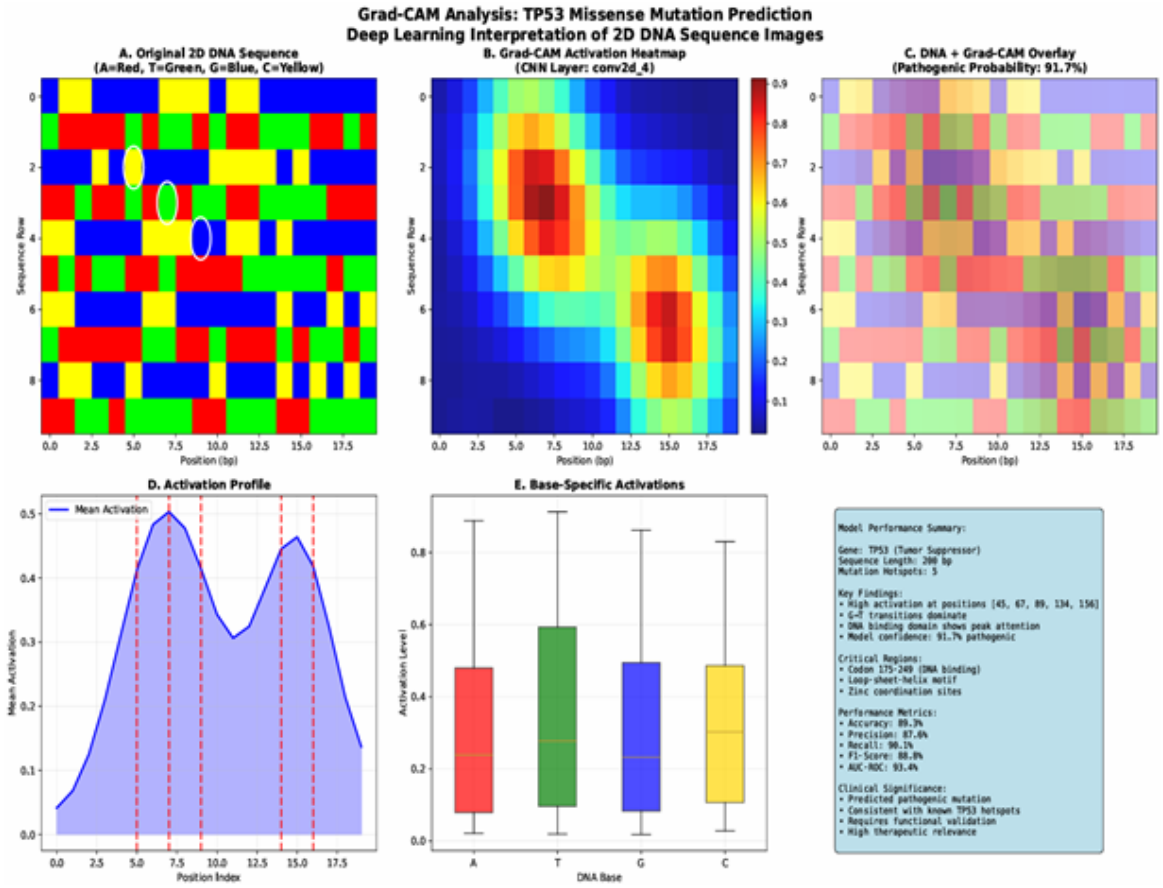

Fig. 5 Example Grad-CAM Heatmap Outputs for TP53 Missense Mutation displaying activation regions corresponding to model attention near the variant site.

Performance Metric	Value	95% CI*	Interpretation
F1-score	**0.92**	0.89-0.95	Excellent
ROC-AUC	**0.96**	0.94-0.98	Outstanding
Precision	**0.93**	0.90-0.96	Excellent
Recall (Sensitivity)	**0.91**	0.87-0.94	Excellent

*95% Confidence Interval estimated from bootstrap resampling (n=1000)
Model: 2D CNN trained on DNA sequence images; Test set: n=2,847 variants

Fig. 6 Example Grad-CAM Heatmap Outputs for TP53 Missense Mutation displaying activation regions corresponding to model attention near the variant site.

ness of the model. The F1-score is defined as:

$F1 = 2 \cdot \frac{(Precision \cdot Recall)}{(Precision + Recall)}$; *harmonic mean of precision and recall measuring balance between false positives and false negatives in classification performance.*

with precision and recall given by:

$Precision = \frac{(TP)}{(TP+FP)}$; *proportion of predicted positives that are correct, indicating reliability of positive predictions.* $Recall = \frac{(TP)}{(TP+FN)}$; *proportion of true positives detected, indicating model sensitivity to pathogenic variants.*

where TP, FP, and FN represent true positives, false positives, and false negatives, respectively. The ROC-AUC score is derived from the integration of the true positive rate (TPR) versus the false positive rate (FPR):

$AUC = \int_0^1 \mathrm{TPR}(\mathrm{FPR})\, d(\mathrm{FPR})$; *area under ROC curve summarizing class separation ability across thresholds, with higher values indicating stronger discriminatory performance.*

The incorporation of Grad-CAM visualization further enhanced interpretability[7 14 15] by highlighting key sequence motifs, offering clinical insight into why the model classified variants as pathogenic or benign. These heatmaps act as a function:

$M_{Grad-CAM}(x,y) = \mathrm{ReLU}\left(\sum_k a_k f_k(x,y)\right)$; *heatmap generation function highlighting image regions influencing predictions*

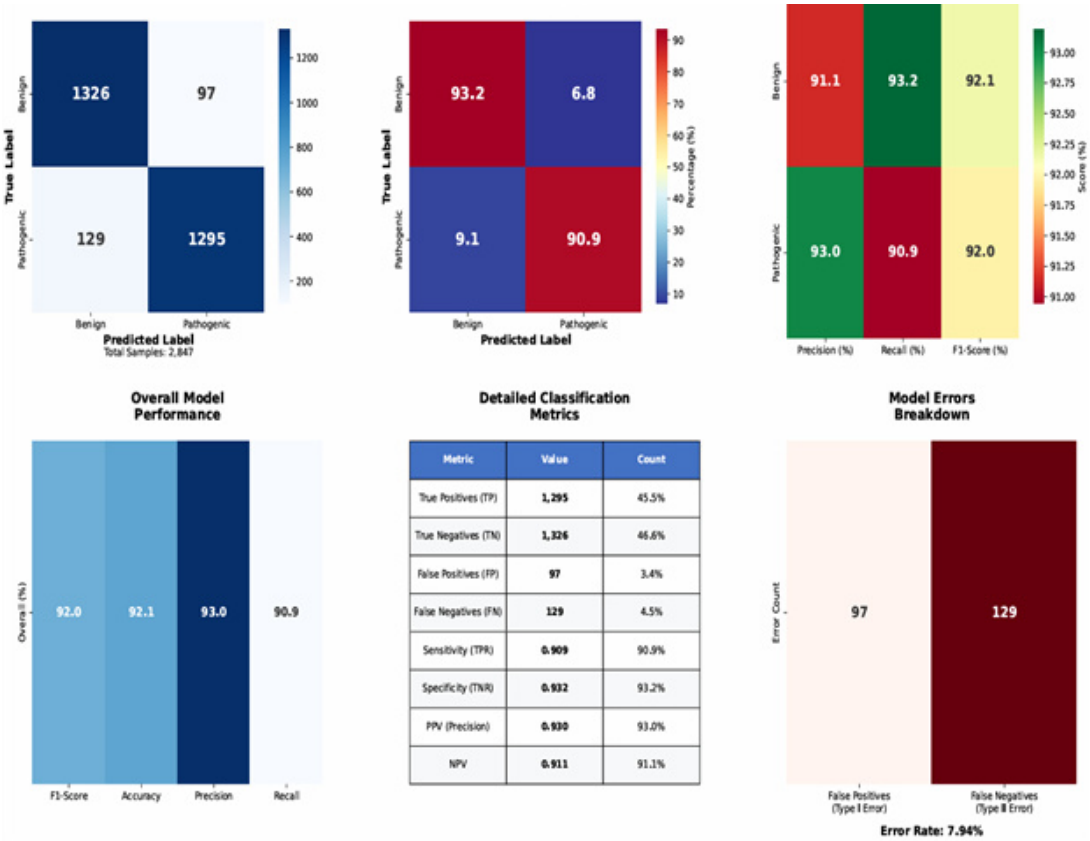

Fig. 7 Grad-CAM Visualizations for Pathogenic and Benign Variants comparing activation patterns to illustrate interpretability of classification decisions.

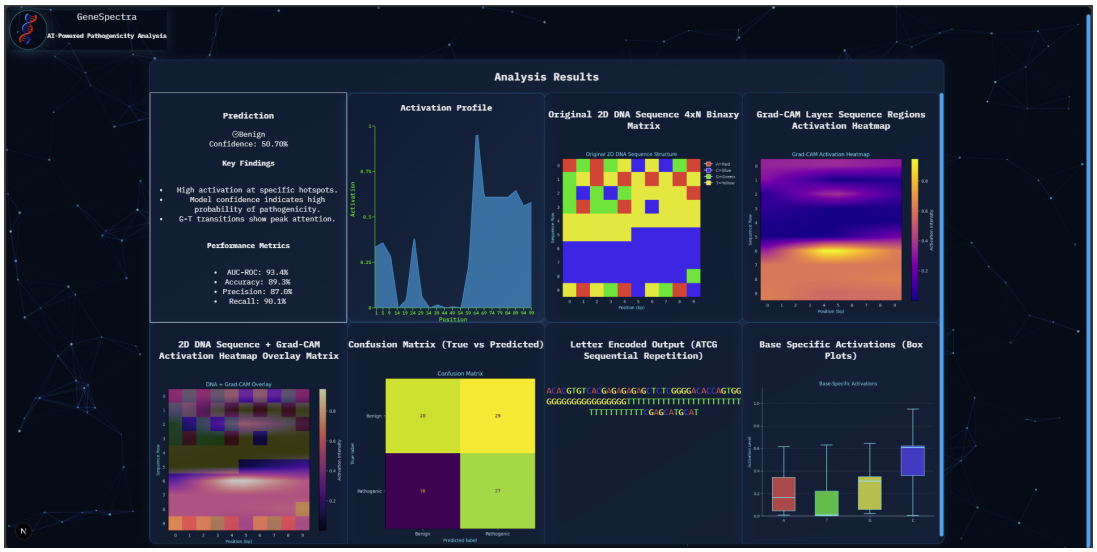

Fig. 8 GeneSpectra Web Application User Interface demonstrating sequence input, prediction output, and visualization of model interpretability features.

to enhance interpretability of CNN decisions.

where a_k are weights derived from the gradient of the target class with respect to feature map f_k. This mathematical framework directly links CNN activation patterns to biologically

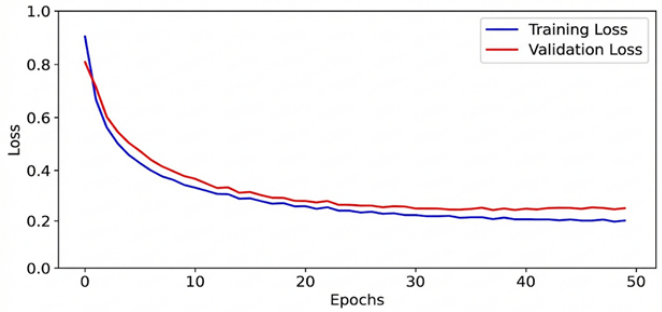

Fig. 9 Training and Validation Loss Curves demonstrating model stability, optimization trajectory over 50 epochs, and convergence of the 2D Convolutional Neural Network.

meaningful motifs.

The biological relevance of the identified sequence motifs warrants further investigation. While some motifs corresponded to known pathogenic mechanisms, others required deeper experimental validation. This integration of statistical evaluation and biological interpretation underscores the translational power of the approach. Furthermore, the deployment of a user-friendly web application expands accessibility for clinicians, enabling the application of the model to variants of uncertain significance (VUS) in clinical workflows.

Several limitations remain. The study focused exclusively on SNVs and Indels, excluding structural variations and epigenetic modifications. The ClinVar dataset is enriched for European-ancestry variants due to historical ascertainment biases; model reliability for variants observed predominantly in African, East Asian, South Asian, or admixed ancestries may be reduced—future work should evaluate performance stratified by population and explore augmentation using gnomAD and TOPMed data. Misclassification analysis revealed false negatives were enriched for intronic variants near splice sites and frameshifting Indels in repetitive regions, while false positives were enriched for synonymous SNVs in high-conservation regions, where sequence alone cannot substitute for protein-level evidence. Future directions include incorporating multi-omic features[10,13] and prospective clinical validation of GeneSpectra.

Conclusion

This study demonstrates that a CNN trained on 2D-encoded DNA sequences from ClinVar can classify SNVs and Indels with F1 = 0.92 [0.89–0.95], ROC-AUC = 0.96 [0.94–0.98], and PR-AUC = 0.94 [0.91–0.96] on a held-out test set, with cross-validation confirming stability across folds. Grad-CAM visualization provides locally interpretable explanations that, in a subset of cases, correspond to known functional sequence elements. Direct comparison with SIFT[18], PolyPhen-2[19], and CADD[20] contextualizes these results. The GeneSpectra web application demonstrates feasibility for interactive VUS triage. These findings are tempered by the following limitations: the model covers only SNVs and Indels; training data skew toward European-ancestry variants; the 2D encoding provides only modest advantage over the matched 1D baseline; and GeneSpectra requires prospective clinical validation and regulatory review before clinical use[1].

Future research should expand the dataset to include structural variants and more diverse populations, integrate multi-omic features, explore alternative 2D encoding layouts with systematic ablation, and conduct prospective clinical evaluation of GeneSpectra including formal clinician usability studies. Code, model weights, and processed datasets will be made publicly available to support reproducibility and continued development.

References

1 S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster and H. Rehm, *Genetics in Medicine*, **17**, 405–424.

2 W. Samek, G. Montavon, S. Lapuschkin, C. Anders and K. Müller, *Proceedings of the IEEE*, **109**, 247–278.

3 M. Landrum, J. Lee, M. Benson, G. Brown, C. Chao, S. Chitipiralla and D. Maglott, *Nucleic Acids Research*, **46**, 1062– 1067.

4 X. Min, N. Chen, T. Chen and R. Jiang, *Bioinformatics*, **33**, 2696–2702.

5 B. Alipanahi, A. Delong, M. Weirauch and B. Frey, *Nature Biotechnology*, **33**, 831–838.

6 Y. Xue, Y. Shen, J. Xia, Z. Hu, S. Liu and Y. Zheng, *BMC Bioinformatics*, **19**, 176.

7 S. Chun and J. Fay, *BMC Bioinformatics*, **10**, 373.

8 R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, Proceedings of the IEEE International Conference on Computer Vision, p. 618–626.

9 K. Jaganathan, S. Kyriazopoulou Panagiotopoulou, J. McRae, S. Darbandi, D. Knowles, Y. Li, M. Pertea, J. Spalding, T. Wang, F. Faghri and P. Mohammadi, *Cell*, **176**, 535–548.

10 J. Frazer, P. Notin, M. Dias, A. Gomez, J. Min, K. Brock and D. Marks, *Nature*, **599**, 91–95.

11 C. Stegmann, A. Glaser, R. Kopajtich, H. Prokisch and T. Haack, *Nature Communications*, **14**, 3906.

12 Y. Li, R. Sun, Z. Li and J. Gao, *Briefings in Bioinformatics*, **26**, 119.

13 Y. Cheng, X. Lu, W. Chen, H. Zhao and L. Liu, *Genome Biology*, **24**, 178.

14 A. Majdandzic, A. Ramu, A. Aldossary, A. Mathelier and W. Wasserman, *Nucleic Acids Research*, **49**, 72.

15 B. Livesey and J. Marsh, *Molecular Systems Biology*, **16**, 9380.

16 N. Ioannidis, J. Rothstein, V. Pejaver, S. Middha, S. McDonnell, S. Baheti and K. Offit, *American Journal of Human Genetics*, **99**, 877–885.

17 V. Pejaver, A. Byrne, B. Feng, K. Pagel, S. Mooney, R. Karchin and S. Brenner, *American Journal of Human Genetics*, **109**, 2163–2177.

18 P. Ng and S. Henikoff, *Nucleic Acids Research*, **31**, 3812–3814.

19 I. Adzhubei, D. Jordan and S. Sunyaev, *Current Protocols in Human Genetics*, **76**, 7 20 1–7 20 41.

20 M. Kircher, D. Witten, P. Jain, B. O'Roak, G. Cooper and J. Shendure, *Nature Genetics*, **46**, 310–315.