# Computational Approaches for Gene Detection: A Comparative Analysis

Mohammad Al Bataineh*
*College of Engineering, Electrical and Communication Engineering Department*
*United Arab Emirates University*
Al Ain, United Arab Emirates
mffbataineh@uaeu.ac.ae
*Corresponding author

Dana I. Abu-Abdoun
*College of Engineering, Electrical and Communication Engineering Department*
*United Arab Emirates University*
Al Ain, United Arab Emirates
dideisan@uaeu.ac.ae

Sameh Al-Shihabi
*Industrial Engineering and Engineering Management*
*University of Sharjah*
Sharjah, United Arab Emirates
salshihabi@sharjah.ac.ae

Khalid Muhammad
*College of Science, Biology Department*
*United Arab Emirates University*
Al Ain, United Arab Emirates
k.muhammad@uaeu.ac.ae

***Abstract*— This study presents a comparative analysis of various computational approaches used for gene detection, emphasizing their efficiency and the potential benefits of integrating multiple methodologies. The paper examines digital signal processing (DSP) techniques, hidden Markov models (HMMs), machine learning-based gene detection algorithms, and gene-finding software to determine which methods yield the most accurate results. Also, we explore whether a hybrid integration of these approaches could enhance gene detection outcomes. A critical review of existing literature suggests that while machine learning models, particularly deep learning frameworks, demonstrate superior predictive performance, DSP and HMM techniques retain their relevance due to their strong theoretical underpinnings. The integration of multiple methods, leveraging their complementary strengths, holds promise for advancing the field of computational gene detection.**

***Keywords—Gene expression, Gene detection, Digital signal processing, Genomic signal processing, Machine learning algorithms.***

## I. Introduction

The identification of genes within a genome is a fundamental challenge in bioinformatics, critical for understanding genetic functions, diagnosing genetic disorders, and advancing fields such as precision medicine and synthetic biology [1]. Genes are functional segments of DNA that encode proteins, and their accurate identification is essential for elucidating biological mechanisms at the molecular level [2]. However, gene detection remains complex due to the inherent variability in genomic sequences, the presence of noncoding regions, and the diversity of genetic structures across species [3]. In prokaryotes, genes typically exist as continuous open reading frames (ORFs), whereas in eukaryotes, they are fragmented into exons separated by noncoding introns, making computational approaches indispensable for their accurate identification [4].

Over the years, a variety of computational methodologies have been developed for gene detection, ranging from rule-based statistical models to advanced machine-learning techniques [5]. Among the prominent approaches are Digital Signal Processing (DSP) techniques, which analyze periodicity in DNA sequences to identify coding regions; Hidden Markov Models (HMMs), which employ probabilistic models to predict gene structures; machine learning and deep learning algorithms, which use data-driven methods to classify genomic sequences; and gene-finding software, which integrates multiple computational strategies to enhance accuracy [6, 7]. While each approach has demonstrated success in specific contexts, there is no universally superior method, necessitating comparative evaluations and potential integrations for improved performance.

One of the primary challenges in gene detection is distinguishing between coding and noncoding regions, especially in eukaryotic genomes where intronic sequences interrupt exons [8]. The presence of alternative splicing further complicates this task, as a single gene can give rise to multiple transcripts with different exon-intron compositions [5]. Moreover, genomic sequences exhibit species-specific characteristics, meaning that models trained on one organism may not generalize well to others, requiring adaptation or retraining for diverse datasets [10].

From a computational perspective, accuracy, scalability, and interpretability are critical considerations in gene detection methodologies. Rule-based statistical models such as HMMs provide interpretable results but may lack adaptability, while deep learning models offer high predictive accuracy but often function as "black boxes," limiting their biological interpretability [11]. Furthermore, the vast size of genomic datasets necessitates efficient algorithms that can process millions of base pairs within reasonable time constraints, highlighting the need for computationally scalable solutions [12].

This paper presents a comparative analysis of computational gene detection methodologies, evaluating their efficiency, strengths, and limitations. Specifically, we aim to (1) assess the performance of DSP techniques, HMMs, machine learning algorithms, and gene-finding software in detecting protein-coding genes; (2) identify key challenges associated with each method and discuss potential improvements through hybrid or integrative models; (3) explore future research directions that could enhance gene detection accuracy and scalability, particularly through interdisciplinary techniques such as explainable artificial intelligence and reinforcement learning.

The remainder of the paper is structured as follows: Section II provides a conceptual overview of gene detection methods, outlining the theoretical foundations and key computational approaches. Section III presents a comparative analysis of these methodologies, evaluating their performance across various criteria such as sensitivity, specificity, computational complexity, scalability, and interpretability. Section IV discusses the challenges and limitations of current approaches, as well as potential avenues for future research. Finally, Section V concludes the paper by summarizing key findings and highlighting the implications of integrating multiple computational strategies for gene detection.

## II. Conceptual overview of gene detection methods

Gene detection has seen substantial progress with advancements in computational methodologies, enabling more precise identification of protein-coding regions within DNA sequences. These methodologies fall into four primary categories: digital signal processing (DSP) techniques, hidden Markov models (HMMs), machine learning-based approaches, and gene-finding software. Each of these methods has played a crucial role in addressing challenges related to distinguishing coding from noncoding regions, particularly in complex eukaryotic genomes where exons and introns are interspersed [1], [5].

### *A. Digital Signal Processing (DSP) Techniques*

Digital signal processing techniques have been applied in genomics to identify periodic patterns within DNA sequences, leveraging mathematical transformations to detect key frequency components indicative of protein-coding regions. One of the primary motivations for using DSP methods is the recognition of a period-3 behavior in protein-coding regions, a phenomenon resulting from codon bias during translation [13]. Among DSP techniques, the discrete Fourier transform (DFT) has been widely utilized to convert nucleotide sequences into the frequency domain, allowing the identification of coding regions based on spectral peaks at frequency 1/3 [2]. The DFT for binary sequences $u_A$, $u_T$, $u_C$, and $u_G$ are represented as $U_A$, $U_T$, $U_C$, and $U_G$. The DFT for the numerical series $u_x$ of length N is defined as follows:

$$U_x(k) = \sum_{n=0}^{N-1} u_x(n) e^{i(2\pi kn/N)}, \qquad (1)$$

The DFT power spectrum for signal $u_x$ at frequency k is detailed in Equation 2, where $U[k]$ is the k-th DFT coefficient, summing squared magnitudes of the DFT coefficients for sequences A, T, C, and G.

$$PS(k) \sum_{x\in[A,T,C,G]} | U_x(k)|^2, k = 0, 1, 2, \dots, N-1, \qquad (2)$$

In addition to DFT, discrete wavelet transform (DWT) has been used for multi-resolution analysis of genomic sequences, making it highly effective in analyzing nonstationary genomic signals [13]. Unlike DFT, which captures global frequency components, DWT allows for localized frequency detection, making it particularly useful in genomes where the periodicity of coding sequences varies across different regions [14]. DWT linearly decomposes a signal $x(t)$ as shown in Equation 3, with wavelet coefficients $a_{j,k}$ derived from Equation 4 and wavelet functions $\psi_{j,k}$.

$$x(t) = \sum_k \ \sum_j a_{j,k} \, \psi_{j,k}(t), \qquad (3)$$

$$a_{j,k} =< x(t)\psi_{j,k}(t) >= \int_{-\infty}^{+\infty} x(t)\psi_{j,k}(t)dt, \qquad (4)$$

The signal $x[n]$ is decomposed into a wavelet basis of orthonormal wavelets and scaling functions $\psi_{j,k}$ [20]. These functions, part of various wavelet families, arise from dilations and translations of a foundational "mother" wavelet, as detailed in Equation 5.

$$\psi_{j,k}(t) = \sum^{-\frac{j}{2}} \psi(2^{-j}t - k), j, k \in \mathrm{Z}, \qquad (5)$$

Digital filtering techniques, such as finite impulse response (FIR) and infinite impulse response (IIR) filters, have also been employed to isolate spectral components associated with coding sequences while reducing noise from noncoding regions [15].

Despite the effectiveness of DSP methods, their reliance on numerical representation schemes for DNA sequences introduces potential biases that can affect detection accuracy. Moreover, DSP-based methods may struggle with analyzing genomes that exhibit complex exon-intron structures, limiting their ability to generalize across diverse species [11].

### *B. Hidden Markov Models (HMMs)*

Hidden Markov models have been extensively used in computational gene detection due to their ability to model probabilistic dependencies within genomic sequences. These models treat DNA as a sequence of hidden states representing coding and noncoding regions, with transition probabilities determining the likelihood of moving between states [9]. This probabilistic framework enables HMMs to effectively model exon-intron boundaries, which is especially useful for eukaryotic gene prediction where exons are fragmented [8].

The core of an HMM consists of a set of states, each with defined transition probabilities and emission probabilities that model nucleotide sequences. The Viterbi algorithm is

Identify the applicable funding agency here. If none, delete this text box.

commonly used to decode the most probable sequence of states, allowing for the identification of gene structures [6]. Several extensions of traditional HMMs have been introduced to improve their predictive power. Generalized HMMs (GHMMs) enhance sequence classification by incorporating additional biological constraints, while duration HMMs (DHMMs) model exon length distributions to improve prediction accuracy [5].

While HMMs provide an interpretable framework for gene detection, their performance is highly dependent on high-quality training data and carefully tuned parameters. A key limitation of HMMs is their reliance on predefined transition probabilities, which may not generalize well across different genomes. Furthermore, HMMs struggle to capture long-range dependencies, making them less effective in predicting genes with complex alternative splicing patterns [9].

### C. *Machine Learning-based Approaches*

Machine learning has revolutionized gene detection by enabling data-driven classification of coding and noncoding regions based on large genomic datasets. Unlike traditional rule-based methods, machine learning models can automatically extract relevant sequence features and learn complex relationships within genetic data [7].

Supervised learning algorithms such as support vector machines (SVMs), random forests (RFs), and artificial neural networks (ANNs) have been widely applied to gene prediction tasks [10]. These models are trained using labeled genomic data, allowing them to generalize well to new sequences. More recently, deep learning architectures, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have further improved predictive accuracy in gene detection [16]. CNNs are particularly effective in identifying spatial motifs within DNA sequences, while long short-term memory (LSTM) networks capture long-range dependencies, enhancing the identification of exon-intron boundaries [17].

Despite their high accuracy, machine learning models require extensive computational resources and large annotated datasets for training. Another major challenge is the interpretability of deep learning models, which often function as black-box predictors, limiting biological insights into their decision-making processes. Researchers are actively exploring explainable artificial intelligence (XAI) methods to improve the transparency of machine learning-based gene detection models while maintaining high predictive performance [7].

### D. *Gene-Finding Software*

Gene-finding software integrates multiple computational approaches to automate gene identification across various genomes. These tools have been developed to provide robust gene prediction pipelines that combine elements of DSP, HMMs, and machine learning techniques. Among the most widely used gene-finding software are GeneMark, AUGUSTUS, and GlimmerHMM, each of which employs distinct methodologies for genome annotation [5].

GeneMark is based on hidden Markov models and has been successfully applied to both prokaryotic and eukaryotic genomes. AUGUSTUS, a widely used gene predictor, combines statistical modeling with machine learning techniques to refine gene annotations, making it particularly effective for genomes with complex gene structures [3]. GlimmerHMM builds upon traditional HMM-based gene-finding techniques by incorporating neural network classifiers, improving the accuracy of exon-intron boundary detection [5].

The accuracy of gene-finding software is influenced by the quality of training data and algorithmic parameters used for annotation. While these tools provide efficient solutions for large-scale genome analysis, their predictions can vary depending on the genome being analyzed. Furthermore, manual curation and comparative validation are often necessary to ensure robust gene annotations, especially when dealing with newly sequenced organisms [18].

## III. COMPARATIVE ANALYSIS OF GENE DETECTION METHODS

The accuracy and efficiency of gene detection methods depend on multiple factors, including the complexity of the genomic sequences being analyzed, the underlying computational model, and the trade-offs between sensitivity, specificity, and interpretability. Over the years, different computational approaches have been developed to address the challenges of identifying protein-coding regions, particularly in eukaryotic genomes where exon-intron structures complicate gene prediction.

A key aspect of evaluating gene detection methods is understanding their relative performance in terms of sensitivity, specificity, computational complexity, scalability, and interpretability. These factors play a crucial role in determining the suitability of a given method for different genomic applications, ranging from small bacterial genomes to large-scale eukaryotic transcriptome analysis [8].

### A. *Sensitivity and specificity in gene prediction*

Sensitivity and specificity are two primary metrics for evaluating gene detection performance. Sensitivity measures the proportion of correctly identified coding regions, while specificity evaluates the exclusion of noncoding regions.

DSP-based methods, particularly those relying on Fourier and wavelet transforms, have demonstrated effectiveness in identifying periodic coding signals. However, their sensitivity varies significantly depending on the choice of numerical mapping techniques, which affects their ability to capture exon-intron boundaries [2, 19, 20]. Fourier transform-based methods typically achieve sensitivity levels between 65% and 85%, but their specificity is often lower due to the presence of noise from noncoding regions [15, 21, 22].

HMMs provide better specificity compared to DSP methods, particularly for prokaryotic genomes where coding regions tend to be continuous. In eukaryotic genomes, however, HMMs struggle with detecting alternative splicing events, leading to

TABLE I. COMPARATIVE SUMMARY OF GENE DETECTION METHODS

| Method | *Sensitivity* | *Specificity* | *Computational complexity* | *Scalability* | *Interpretability* | *References* |
|---|---|---|---|---|---|---|
| DSP | 65-85% | 57-82% | Low | High | High | [2, 15, 19, 20, 21,22] |
| HMM | 76-90% | 77-93% | Moderate | Moderate | Moderate | [5,6,9, 23,24] |
| Machine learning | 94-97% | 92-95% | High | High | Low | [16, 25, 26, 27, 28] |
| Gene-Finding Software | 80-94% | 48-95% | Moderate | Moderate | Moderate | [5, 8, 9, 21] |

sensitivity values ranging between 76% and 90%, depending on genome complexity [6, 23, 24]. Despite this, HMMs remain widely used in gene annotation due to their structured probabilistic modeling.

Machine learning-based approaches, particularly deep learning models such as CNNs and recurrent neural networks (RNNs), have shown significant improvements in both sensitivity (94-97%) and specificity (92-95%) [16, 25, 26, 27, 28]. These models can learn complex sequence dependencies that traditional rule-based methods fail to capture. However, their accuracy is contingent upon the availability of large, well-annotated training datasets, which can be a limiting factor in less-characterized genomes [10].

Gene-finding software tools such as GeneMark, AUGUSTUS, and GlimmerHMM integrate multiple computational techniques to optimize gene prediction. While these tools often match the sensitivity levels of HMM-based approaches, they exhibit slightly lower specificity, particularly in complex genomes where noncoding RNA regions can be misclassified as coding sequences [5, 8, 9, 21].

### *B. Computational Complexity and Scalability*

Computational efficiency is a key differentiator among gene detection methods. As indicated in Table I, DSP-based approaches are the most efficient, operating with low computational requirements and enabling real-time analysis of genomic sequences [19]. However, their reliance on predefined periodicity assumptions limits their ability to handle genome-wide annotation tasks [20]. HMM-based methods are moderately efficient, requiring more computational power than DSP approaches due to their need for state transition calculations [8]. However, scalability issues arise when applying HMMs to large eukaryotic genomes, as the number of possible state transitions increases exponentially with sequence length [9].

Machine learning methods, particularly deep learning, are the most computationally demanding. CNN and LSTM-based models require significant GPU acceleration, making them impractical for real-time analysis but highly effective for large-scale genomic datasets [7]. These models can efficiently scale when trained on large datasets but require substantial preprocessing, feature extraction, and hyperparameter tuning.

Gene-finding software provides a balance between computational efficiency and accuracy. Tools such as AUGUSTUS and GlimmerHMM have been optimized to handle whole-genome annotations, but their execution time depends on the dataset size, genome complexity, and model configuration [3].

### *C. Interpretability and Biological Relevance*

One of the critical concerns in computational gene detection is the interpretability of results. DSP methods offer high interpretability, as they provide spectral visualizations that allow researchers to analyze periodic coding signals in a genome [6]. However, their application is limited by the fact that not all coding regions exhibit strict periodicity, particularly in eukaryotic genomes with alternative splicing events.

HMMs provide interpretable results by explicitly modeling sequence dependencies. Researchers can analyze the transition probabilities between coding and noncoding states, making these models valuable for biological interpretation [9]. However, their dependence on predefined training datasets introduces biases, as performance varies depending on the genome used to train the model. Machine learning models, while highly accurate, suffer from low interpretability. CNNs and LSTMs function as black-box models, meaning their decision-making processes are challenging to explain in biological terms [7]. Efforts are underway to develop explainable AI techniques that improve interpretability while maintaining predictive performance [10].

Gene-finding software offers the most biologically interpretable results, as it integrates sequence homology, probabilistic modeling, and machine learning techniques into comprehensive genome annotation pipelines. However, these tools often require manual curation to validate gene predictions, particularly in newly sequenced genomes where training data are limited [3].

The findings in Table 1 indicate that while machine learning-based approaches achieve the highest predictive accuracy, they remain computationally expensive and require improvements in interpretability. DSP-based methods provide rapid and interpretable spectral analysis, but their limited adaptability affects their applicability for large-scale genome annotations.

HMMs and gene-finding software tools provide a balanced trade-off between accuracy and efficiency, making them well-suited for automated genome annotation.

Future research should focus on hybrid approaches that integrate the strengths of multiple methodologies. For instance, combining DSP-based preprocessing with machine learning classifiers could improve accuracy while reducing computational costs. Furthermore, explainable AI techniques in deep learning should be developed to improve model transparency, making them more useful for biological applications. The application of reinforcement learning for dynamic optimization of gene detection models is another promising avenue for future research.

## IV. Discussion and future directions

The comparative analysis of gene detection methods highlights the strengths and limitations of various computational approaches. Machine learning models, particularly deep learning-based techniques, have demonstrated the highest sensitivity (94-97%) and specificity (92-95%), making them the most effective in identifying protein-coding regions. However, their application is constrained by high computational complexity and the lack of interpretability. HMM-based models provide a structured probabilistic framework with moderate accuracy (76-90% sensitivity, 77-93% specificity), making them particularly effective for genome annotation but less capable of handling complex exon-intron structures. DSP techniques, despite their efficiency and interpretability, struggle with complex genome architectures, limiting their sensitivity (65-85%) due to their reliance on periodicity detection. Gene-finding software provides a balance between computational efficiency and accuracy but requires manual tuning to optimize performance for different genome types.

Despite advancements in computational gene detection, several challenges remain unresolved. One of the main challenges is the scalability of deep learning methods. While deep learning achieves the highest accuracy, its computational requirements hinder its application in large-scale genomic studies [7]. Training deep models requires large datasets, extensive hyperparameter tuning, and significant GPU resources, making them less accessible for smaller research groups or real-time applications.

Another major limitation is the interpretability of machine learning models. Most deep learning-based methods function as black-box models, limiting their biological interpretability. Current research in explainable AI (XAI) aims to address this limitation, but practical implementations for gene detection remain in the early stages. Handling alternative splicing events is another challenge, as HMMs and traditional gene-finding tools struggle with detecting alternative splicing, a critical aspect of eukaryotic genome complexity. New models that incorporate graph-based approaches or splicing-aware neural networks are required [31].

Hybrid approaches for efficiency are an area of potential improvement. While machine learning is effective for gene prediction, hybrid models combining DSP-based preprocessing with machine learning classifiers could improve computational efficiency while maintaining high accuracy [32]. Another important challenge is genome variability across species. Most computational models are trained on specific genome types (e.g., human, bacterial) and do not generalize well to newly sequenced species. A more flexible, transfer learning-based approach is needed to adapt models dynamically to diverse genomes [33].

To address these challenges, future studies should explore the development of explainable AI for gene detection. Developing interpretable machine learning models that allow researchers to understand why a particular gene was predicted as coding or noncoding is crucial. Feature attribution techniques such as SHAP (Shapley Additive Explanations) and Grad-CAM (Gradient-weighted Class Activation Mapping) could help make deep learning decisions more transparent [34].

Another promising direction is reinforcement learning for adaptive gene prediction. Unlike traditional models, reinforcement learning (RL) can dynamically adapt to new genomic data, allowing self-improving models for gene annotation in less-characterized genomes [35]. Integrating DSP and AI for hybrid gene detection could also be a significant breakthrough. A potential approach involves combining DSP for feature extraction with deep learning for classification, leveraging the efficiency of DSP while benefiting from the adaptability of machine learning models [36].

Scalable cloud-based genomic analysis can help democratize access to computationally expensive deep learning models, making them feasible for widespread use in gene annotation projects [37]. Finally, enhancing gene-finding software with AI is another promising avenue. Traditional gene-finding tools like GeneMark and AUGUSTUS can be augmented with neural networks to improve annotation accuracy, particularly for non-model organisms where reference datasets are limited.

## V. Conclusion

This review underscores the rapid evolution of computational gene detection methodologies, with deep learning emerging as the most accurate but computationally demanding approach. HMMs and gene-finding software remain widely used for genome annotation, while DSP-based methods offer efficient but less flexible solutions. Future research should focus on developing hybrid models that integrate DSP, probabilistic frameworks, and deep learning for enhanced efficiency and accuracy. Moreover, increasing interpretability through explainable AI and leveraging cloud-based high-performance computing will be critical in advancing genomic research. The integration of reinforcement learning and adaptive AI-driven annotation systems will likely shape the next generation of computational gene detection tools.

## References

[1] N. Goel, S. Singh, and T. C. Aseri, "A comparative analysis of soft computing techniques for gene prediction," *Analytical Biochemistry*, vol. 438, no. 1, pp. 14–21, 2013.

[2] Y. Zhou, L.-Q. Zhou, Z.-G. Yu, and V. Anh, "Distinguish coding and noncoding sequences in a complete genome using Fourier transform," in *Third International Conference on Natural Computation (ICNC 2007)*, vol. 2. IEEE, 2007, pp. 295–299.

[3] M. Stanke and S. Waack, "Gene prediction with a hidden Markov model and a new intron submodel," *Bioinformatics*, vol. 19, no. suppl_2, pp. ii215–ii225, 2004.

[4] J. Liu, J. Gough, and B. Rost, "Distinguishing protein-coding from noncoding RNAs through support vector machines," *PLoS Genetics*, vol. 2, no. 4, p. e29, 2006.

[5] S. Tang, A. Lomsadze, and M. Borodovsky, "Identification of protein coding regions in RNA transcripts," *Nucleic Acids Research*, vol. 43, no. 12, pp. e78–e78, 2015.

[6] S. Winters-Hilt and C. Baribault, "A metastate HMM with application to gene structure identification in eukaryotes," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 1–18, 2010.

[7] Z. Shen, W. Bao, and D.-S. Huang, "Recurrent neural network for predicting transcription factor binding sites," *Scientific Reports*, vol. 8, no. 1, p. 15270, 2018.

[8] M. Rho, H. Tang, and Y. Ye, "FragGeneScan: Predicting genes in short and error-prone reads," *Nucleic Acids Research*, vol. 38, no. 20, pp. e191–e191, 2010.

[9] K.-J. Won, A. Pru¨gel-Bennett, and A. Krogh, "Training hmm structure with genetic algorithm for biological sequence analysis," *Bioinformatics*, vol. 20, no. 18, pp. 3613–3619, 2004.

[10] J. Pérez-Rodríguez and N. García-Pedrajas, "An evolutionary algorithm for gene structure prediction," in Modern Approaches in Applied Intelligence: 24th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2011, Syracuse, NY, USA, June 28–July 1, 2011, Proceedings, Part II 24. Springer, 2011, pp. 386–395.

[11] H. M. Wassfy, M. L. Salem, M. M. Abdelnaby, M. S. Mabrouk, and A. Zidan, "Advanced DNA mapping schemes for exon prediction using digital filters," *American Journal of Biomedical Engineering*, vol. 6, no. 1, pp. 25–31, 2016.

[12] M. Al Bataineh and Z. Al-qudah, "A novel gene identification algorithm with Bayesian classification," *Biomedical Signal Processing and Control*, vol. 31, pp. 6–15, 2017.

[13] T. P. George and T. Thomas, "Discrete wavelet transform de-noising in eukaryotic gene splicing," *BMC Bioinformatics*, vol. 11, pp. 1–8, 2010.

[14] O. Abbasi, A. Rostami, and G. Karimian, "Identification of exonic regions in DNA sequences using cross-correlation and noise suppression by discrete wavelet transform," *BMC Bioinformatics*, vol. 12, pp. 1–10, 2011.

[15] M. K. Hota and V. K. Srivastava, "Performance analysis of different DNA to numerical mapping techniques for identification of protein coding regions using tapered window based short-time discrete Fourier transform," in *2010 International Conference on Power, Control and Embedded Systems*. IEEE, 2010, pp. 1–4.

[16] Y. Zhang, S. Qiao, S. Ji, and Y. Li, "DeepSite: Bidirectional LSTM and CNN models for predicting DNA–protein binding," *International Journal of Machine Learning and Cybernetics*, vol. 11, pp. 841–851, 2020.

[17] N. Q. K. Le, D. T. Do, T. N. K. Hung, L. H. T. Lam, T.-T. Huynh, and N. T. K. Nguyen, "A computational framework based on ensemble deep neural networks for essential genes identification," *International Journal of Molecular Sciences*, vol. 21, no. 23, p. 9070, 2020.

[18] S. D. Sharma, K. Shakya, and S. Sharma, "Evaluation of DNA mapping schemes for exon detection," in *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)*. IEEE, 2011, pp. 71–74.

[19] H. Saberkari, M. Shamsi, H. Heravi, and M. H. Sedaaghi, "A fast algorithm for exonic regions prediction in dna sequences," *Journal of medical signals and sensors*, vol. 3, no. 3, p. 139, 2013.

[20] M. R. Kumar and N. K. Vaegae, "A new numerical approach for dna representation using modified gabor wavelet transform for the identification of protein coding regions," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 2, pp. 836–848, 2020.

[21] M. Al Bataineh, "Identification of coding regions in prokaryotic dna sequences using bayesian classification," in Bioinformatics and Biomedical Engineering: 8th International Work-Conference, IWBBIO 2020, Granada, Spain, May 6–8, 2020, Proceedings 8. Springer, 2020, pp. 3–14.

[22] N. Q. K. Le, D. T. Do, T. N. K. Hung, L. H. T. Lam, T.-T. Huynh, and N. T. K. Nguyen, "A computational framework based on ensemble deep neural networks for essential genes identification," *International journal of molecular sciences*, vol. 21, no. 23, p. 9070, 2020.

[23] S. Winters-Hilt, Z. Jiang, and C. Baribault, "Hidden markov model with duration side information for novel hmmd derivation, with application to eukaryotic gene finding," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 1–11, 2010.

[24] A. M. Khedr, M. H. Ibrahim, and A. Al Ali, "Lpb: a new decoding algorithm for improving the performance of an hmm in gene finding application," *IAENG International Journal of Computer Science*, vol. 47, no. 4, pp. 723–729, 2020.

[25] A. Al-Ajlan and A. El Allali, "Cnn-mgp: convolutional neural networks for metagenomics gene prediction," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 11, pp. 628–635, 2019.

[26] R. Silva, K. Padovani, F. Góes, and R. Alves, "GeneRFinder: Gene finding in distinct metagenomic data complexities," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–17, 2021.

[27] M. Emam, A. Ali, E. Abdelrazik, M. Elattar, and M. El-Hadidi, "Detection of mammalian coding sequences using a hybrid approach of chaos game representation and machine learning," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 2949–2951.

[28] C. Wei, J. Zhang, and X. Yuan, "Enhancing the prediction of protein coding regions in biological sequence via a deep learning framework with hybrid encoding," *Digital Signal Processing*, vol. 123, p. 103430, 2022.

[29] Q. Zheng, T. Chen, W. Zhou, S. A. Marhon, L. Xie, and H. Su, "SAVMD: An adaptive signal processing method for identifying protein coding regions," *Biomedical Signal Processing and Control*, vol. 70, p. 102998, 2021.

[30] L. Das, S. Nanda, and J. Das, "An integrated approach for identification of exon locations using recursive Gauss Newton tuned adaptive Kaiser window," *Genomics*, vol. 111, no. 3, pp. 284–296, 2019.

[31] I. Provazník, V. Kubicová, H. Škutková, J. Nedvěd, E. Tkacz, P. Babula, and R. Kizek, "Detection of short exons in DNA sequences using complex wavelet transform of structural features," in *Proceedings 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*. IEEE, 2012, pp. 107–110.

[32] X. Zhang, J. Zhao, and W. Xu, "Identification of eukaryotic exons using empirical mode decomposition and modified Gabor-wavelet transform," in *Proceedings of the 33rd Chinese Control Conference*. IEEE, 2014, pp. 7151–7155.

[33] V. Pathak, S. Jagannath Nanda, A. Mahesh Joshi, and S. Sekhar Sahu, "Hardware implementation of infinite impulse response anti-notch filter for exon region identification in eukaryotic genes," *International Journal of Circuit Theory and Applications*, vol. 48, no. 12, pp. 2242–2256, 2020.

[34] Q. Zhang and Y. Jiang, "An abnormal gene detection method based on Selene," in Intelligent Computing Theories and Application: 17th International Conference, ICIC 2021, Shenzhen, China, August 12–15, 2021, Proceedings, Part III 17. Springer, 2021, pp. 396–406.

[35] M. Emam, A. Ali, E. Abdelrazik, M. Elattar, and M. El-Hadidi, "Detection of mammalian coding sequences using a hybrid approach of chaos game representation and machine learning," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2020, pp. 2949–2951.

[36] C. Wei, J. Zhang, and X. Yuan, "Enhancing the prediction of protein coding regions in biological sequence via a deep learning framework with hybrid encoding," *Digital Signal Processing*, vol. 123, p. 103430, 2022.

[37] Y. Wang, X. Dai, D. Fu, P. Li, and B. Du, "PGD: A machine learning-based photosynthetic-related gene detection approach," *BMC Bioinformatics*, vol. 23, no. 1, p. 183, 2022.

AUTHORS' INFORMATION

| **Your Name** | **Title*** | **Research Field** | **Personal website** |
|---|---|---|---|
| Mohammad Al Bataineh | Assistant Professor | Signal processing, coding and information theory, bioinformatics | https://scholar.google.co.uk/citations?user=vXVMff8AAAAJ&hl=en&oi=ao |
| Dana I. Abu-Abdoun | Research assistant | Data mining, machine learning algorithms | https://scholar.google.co.uk/citations?user=XQCmHKUAAAAJ&hl=en&oi=ao |
| Sameh Al Shihabi | Professor | Operations research | https://scholar.google.co.uk/citations?user=NRBuaZ0AAAAJ&hl=en&oi=ao |
| Khalid Muhammad | Associate Professor | Immunology, molecular biology | https://scholar.google.co.uk/citations?user=uJD33iwAAAAJ&hl=en&oi=ao |