International Journal of Research in Engineering, Science and Management
Volume 8, Issue 5, May 2025
https://www.ijresm.com | ISSN (Online): 2581-5792

Secure Medical Diagnostics via Digital DNA Sequencing and Machine Learning

Sushalini Ravi[1*], Daphney Joann[2], Kayalvizhi Loganathan[3], Harinie Sekar[4]
[1,3,4]Student, Department of Computer Science and Engineering, Global Institute of Engineering and Technology, Ranipet, India
[2]Assistant Professor, Department of Artificial Intelligence & Data Science, Global Institute of Engineering & Technology, Ranipet, India

***Abstract*: This paper presents a cutting-edge web-based healthcare management system that offers safe and effective healthcare services while bridging the gap between hospitals, physicians, and patients. Hospitals can use the platform to manage physicians, register patients, and set service fees. Convenient hospital and doctor selection, safe payment processing, and a personalized token that allows encrypted contact with the selected physician are all available to patients. Patients and physicians may communicate easily thanks to the system's encrypted chat interface. Doctors use the Smith-Waterman method to examine DNA sequences uploaded by patients, producing accurate diagnosis results. The platform is managed by an administrator, who guarantees efficient operations and strong security compliance. The system uses the AES encryption method to safeguard private patient data, guaranteeing data integrity and confidentiality throughout transmission and storage. The technology provides a novel and individualized approach to contemporary healthcare by fusing secure communication channels with cutting-edge DNA analysis capabilities, promoting efficiency, security, and confidence in medical services.**

***Keywords*: AES Encryption, DNA analysis, Healthcare Management, Medical Diagnostics, Personalized Healthcare, Smith-Waterman Algorithm.**

1. Introduction

Technology integration has significantly improved the healthcare sector, resulting in safer and more effective administration of medical services. By providing a smooth and safe platform for communication between hospitals, physicians, and patients, this web-based healthcare management system seeks to improve the overall healthcare experience. Patients can choose hospitals and doctors according to their preferences thanks to its facilitation of hospital registration, doctor administration, and service cost definitions. Patients can engage with physicians in a secure and private setting by implementing token-based authentication and encrypted communication through secure chat. Additionally, the system allows users to input DNA sequences for analysis using the Smith-Waterman algorithm, providing accurate diagnostic results customized for each patient. By using AES encryption, confidentiality and trust are maintained and all patient data is safely kept and transferred. An administrator keeps an eye on the system to make sure everything runs well and security regulations are followed. The goal of this project is to create a safe, creative, and effective healthcare environment while bridging the gap between patients and healthcare providers.

The suggested system is an all-inclusive web-based healthcare platform made to give hospitals, physicians, and patients a productive, safe, and easy-to-use environment. The technology makes it easier for doctors and hospitals to register, giving the latter the ability to manage their physician profiles, establish service fees, and offer patients comprehensive medical care. A novel approach to store RDF graph data utilizing DNA-based storage technologies was presented in the prior paper, indicating a forward-looking path for data analytics [1]. To improve the accuracy of genetic analysis, a sequential labeling model is created to jointly classify mutation types and indices in lung cancer DNA [2]. Their opinions on the gathering, application, and privacy issues of DNA, facial photos, and other biometric information reveal differing degrees of confidence in organizations that manage biometrics [3]. A deep learning-based method for analyzing and identifying harmful DNA sequences that can jeopardize the privacy of genetic data is given. The project intends to solve new issues in bio informatics and cybersecurity by utilizing deep neural networks to improve security in DNA processing and storage systems [4].

A virtual platform for multiscale DNA nanostructure modeling and visualization that facilitates improved DNA nanotechnology analysis and design [5]. In order to improve the dependability and effectiveness of DNA-based data storage systems, it investigates the architecture and properties of restricted codes [6]. By combining a Viterbi decoder with the basecaller and using convolutional codes, it offers a novel way to lessen high error rates in nanopore-based DNA data storage, resulting in up to three times lower reading costs than current methods [7]. In order to accomplish effective and balanced encoding for DNA storage systems, it suggests a DNA data compression technique based on a minimum variance Huffman tree [8]. As a conceptual extension of conventional scale-space methodologies, it presents Scale Trotter, a revolutionary visualization technique that allows users to explore data across negative sizes [9].

Decal-Lenses is a cutting-edge visualization method made for directly examining multivariate data on three-dimensional surfaces. Decal-lenses are affixed to the surface geometry,

*Corresponding author: sushaliniravi1303@gmail.com

functioning as stickers or decals that adapt to the surface contours, in contrast to conventional lenses that hover above the visualization space [10]. Under the present approach, patients can share their DNA sequences with physicians, who will use the Smith-Waterman algorithm to analyze the data and produce a diagnosis. Physicians will be able to communicate with patients via a chat interface, discuss diagnosis, and offer tailored medical recommendations. Administrative controls are also included into the platform to guarantee seamless operation, keep an eye on hospital and physician registration, and enforce security guidelines.

This study offers a thorough method for protecting medical diagnostic systems against ransomware assaults by utilizing machine learning and digital DNA sequencing. The system architecture is presented in section 2, which also describes how the ransomware detection module is included into medical diagnostic procedures without interfering with essential operations. In order to simulate malware activity, section 3 explains the Digital DNA sequencing method. This method involves encoding malicious behaviors into structured sequences that resemble DNA, allowing for more efficient behavior-based analysis. The experimental results are covered in section 4, which also highlights the accuracy, precision, and robustness of the method by demonstrating the detection performance utilizing a variety of machine learning techniques. In order to further increase the security of healthcare systems, section 5 summarizes the contributions and lays out future directions.

2. System Architecture

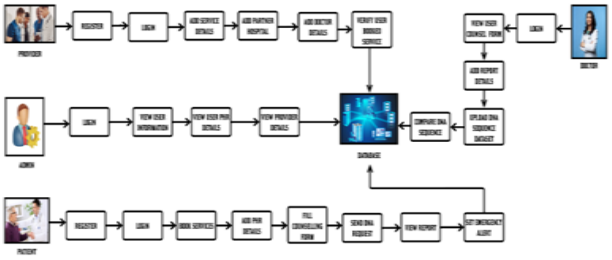

Fig. 1. System architecture of secure medical diagnostics

A web-based healthcare management system that makes it easier for patients, physicians, and hospitals to communicate safely and effectively. Users can choose services according to their interests, as seen in figure 1, and hospitals can register on the site to manage doctor profiles and specify service fees. In addition to making safe payments and receiving a personal token for secure communication, patients can peruse and select hospitals and physicians. The platform uses the Smith-Waterman algorithm to give accurate diagnostic results by enabling patients to share DNA sequences with physicians for study. The paper offers a user-friendly interface for seamless transactions and data sharing, focusing on personalized healthcare services.

A. Digital DNA Sequencing for Malware Behavior Modeling

This module converts malware behavior into structured, DNA-like sequences to facilitate advanced behavior-based analysis. System activities—such as API calls, file operations, registry modifications, or network behaviors—are encoded into symbolic representations similar to genetic nucleotides (e.g., A, T, C, G), where each symbol represents a specific category or pattern of behavior. These encoded sequences mimic the structure of biological DNA, capturing the temporal and semantic nature of malicious operations. By modeling malware in this way, the module enables the detection of behavioral "mutations" or deviations from normal activity, which are indicative of obfuscation or evolving ransomware tactics. The system's capacity to recognize unknown or polymorphic threats is enhanced when the resultant digital DNA sequences are fed into downstream machine learning classifiers, which identify and classify malware based on its behavioral patterns rather than static signatures.

B. Steps in Digital DNA Sequencing Module

1) Behavioral Event Logger

Input: Raw activity logs from monitoring agents or sandboxes.

Output: Structured logs of behavioral events with timestamps and metadata.

2) Symbolic Encoder

Function: Maps behavioral events to nucleotide-like symbols (e.g., A = File Access, T = Registry Change, C = Network Request, G = Process Creation).

Logic: Uses predefined encoding rules or learned mappings based on frequency and context.

Output: Behavior sequences in symbolic form (e.g., ATCGGATC).

3) Sequence Constructor

Function: Assembles encoded symbols into complete DNA-like sequences representing the malware's behavior flow.

Features: Supports time-windowing and segmentation to preserve context and order.

Output: Full-length digital DNA sequences per malware instance.

4) Mutation Pattern Detector

Function: Identifies anomalies in behavior sequences compared to baseline benign patterns.

Goal: Detects behavior "mutations" that may indicate obfuscation, packing, or ransomware activity.

5) K-mer Generator & Feature Extractor

Function: Breaks sequences into fixed-length substrings (k-mers) to capture local behavior patterns.

Purpose: Enables statistical and structural analysis for machine learning input.

Output: Feature vectors representing malware behavior.

6) Sequence Storage & Indexing

Function: Stores generated digital DNA sequences in an optimized database.

Features: Allows rapid retrieval, comparison, and update of sequences.

Use Case: For reanalysis, training, and threat intelligence correlation.

Table 1
Sample DNA sequences with constraints

S.No.	DNA Sequence	GC Content (%)	Melting Temp (°C)	Hairpin Present
1	GTACCTAGTGCCTGTAATGC	50.00	60.00	No
2	ATGCGTATGCCGTACAGTTA	55.00	62.00	No
3	TACGGTAGCTAGGTTACGCA	60.00	64.00	No
4	CGTATGCCTAGCTACGTGCA	55.00	62.00	No
5	AGCTAGCTGATCGTGACGTA	50.00	60.00	No

3. Results and Discussions

A collection of DNA sequences created with physiologically inspired design constraints to guarantee structural stability and accurate hybridization. Each 20 nucleotide sequence is assessed according to three crucial factors: the presence of hairpin structure, melting temperature (Tm), and GC concentration. The amount of guanine (G) and cytosine (C) bases that contribute to the stability of the DNA duplex is represented by the GC content, which is stated as a percentage. In this case, a range of 40% to 60% is appropriate. To ensure dependable hybridization throughout molecular processes, the melting temperature which is determined by the Wallace rule and indicates the temperature at which the DNA strand dissociates was maintained between 50°C and 70°C.Hairpin structures, which are inadvertent intra-strand loops created by palindromic regions and might hinder molecular recognition and decrease binding selectivity, were also avoided by screening sequences. Since all of the sequences on the list adhere to these restrictions, they can be used safely and effectively in applications such as malware behavior encoding, computation, and DNA-based diagnostics.

To ensure secure and private interaction between a patient and a doctor using a unique, encrypted Token ID. The inputs are Patient Token ID, Doctor Session Access and Message or Document Upload.

Fig. 2.

Using distinct token IDs, it ensures a session-bound, encrypted, and traceable communication route between physicians and patients. Through the technology, people can safely provide their DNA sequences to healthcare providers, who can then use a variety of bioinformatics algorithms to gain diagnostic insights. Because of its great accuracy in local sequence alignment and ability to precisely detect regions of similarity between genetic sequences, the Smith-Waterman technique is used among them. The Needleman-Wunsch technique, which yields global alignment findings crucial for identifying overall sequence similarity, is utilized for full-length sequence comparisons. In cases where speed is crucial, especially when scanning large genomic databases, doctors may use BLAST (Basic Local Alignment Search Tool), which offers fast, heuristic-based matching while maintaining reasonable accuracy. Hidden Markov Models (HMMs) are also used to discover complicated patterns, like gene architectures or mutations, that may be suggestive of hereditary disorders and to predict sequence variability. When combined, these algorithms offer a complete toolkit for DNA sequence analysis that strikes a balance between accuracy, speed, and scalability to enable precise and effective diagnosis.

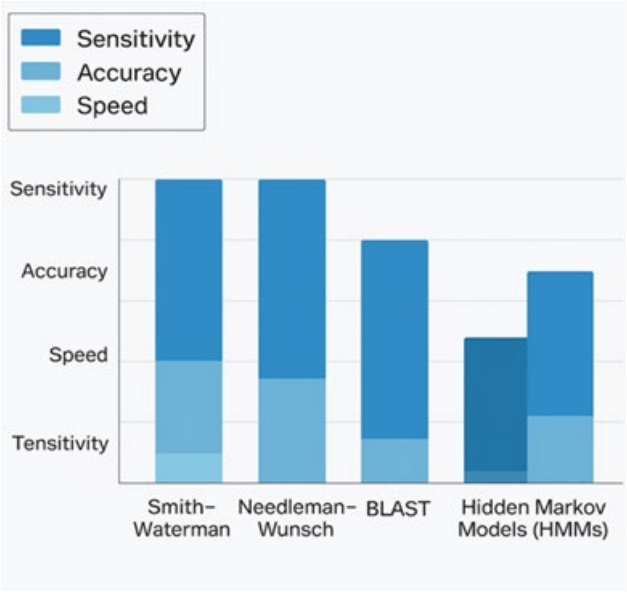

Fig. 3. Comparison chart of various algorithms providing DNA sequencing

The bar graph contrasts the sensitivity, accuracy, and speed of the four main sequence alignment algorithms: BLAST, Needleman-Wunsch, Smith-Waterman, and Hidden Markov Models (HMMs). Because of their dynamic programming methodology, the Smith-Waterman and Needleman-Wunsch algorithms both provide great sensitivity and accuracy; nevertheless, their relative slowness makes them impractical for large-scale investigations. In contrast, BLAST is a speed-oriented algorithm that effectively searches big datasets, but at the expense of some sensitivity and accuracy because of its heuristic character. Protein families and domain structures can be well-modeled by Hidden Markov Models because they offer a balanced solution with high sensitivity and accuracy, particularly for recognizing distant homologs, and moderate speed. Overall, the graph illustrates how various techniques compromise computing efficiency and accuracy.

4. Conclusion

In summary, the suggested web-based healthcare management system provides a practical and effective way to improve communication between patients, physicians, and

hospitals. The system guarantees that patients and physicians can communicate by incorporating elements like communication, DNA sequence analysis using the Smith-Waterman algorithm, and an easy-to-use interface. While the platform's administrative controls guarantee smooth operation and regulatory compliance, the usage of AES encryption offers strong security, safeguarding personal data. By facilitating individualized diagnostic solutions, this system not only raises the standard of healthcare services but also makes it easier for physicians and hospitals to manage their services. Additionally, the system can be upgraded to include future technology because it is scalable. By fusing cutting-edge technology with an emphasis on security and usability, the system claims to boost patient trust, improve healthcare delivery, and build a more effective ecosystem of medical services.

References

[1] Asad Usmani And Lena Wiese, "DNA-Based Storage of RDF Graph Data: A Futuristic Approach to Data Analytics," November 2023.

[2] Untari Novia Wisesty, Ayu Purwarianti, Adi Pancoro, Amrita Chattopadhyay, Nam Nhut Phan, Eric Y. Chuang, And Tati Rajab Mengko, "Join Classifier of Type and Index Mutation on Lung Cancer DNA Using Sequential Labeling Model," January 2022.

[3] Sara H. Katsanis, Peter Claes, Megan Doerr, Robert Cook-Deegan, Jessica D. Tenenbaum, Barbara J. Evans, Myoungkeun Lee, Joel Anderton, Seth M. Weinberg, And Jennifer K. Wagner, "U.S. Adult Perspectives on Facial Images, DNA, and Other Biometrics," March 2022.

[4] Ho Bae, Seonwoo Min, Hyun-Soo Choi, And Sungroh Yoon, "DNA Privacy: Analyzing Malicious DNA Sequences Using Deep Neural Networks," March 2022.

[5] David Ku Tak, Matias Nicolas Selzer, Jan By Ska, Mari´ A Lujan Ganuza, Ivan Bari Sic, Barborakozlikova, and Haichao Miao, "Vivern–A Virtual Environment for Multiscale Visualization and Modeling of DNA Nanostructures.," December 2022.

[6] A. S. Immink and K. Cai, "Properties and constructions of constrained codes for DNA-based data storage," IEEE Access, vol. 8, pp. 49523-49531, 2020.

[7] S. Chandak et al., "Overcoming high nano pore base caller error rates for DNA storage via base caller-decoder integration and convolutional codes," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2020, pp. 8822-8826.

[8] P. Mishra, C. Bhaya, A. K. Pal, And A. K. Singh, "Compressed DNA coding using minimum variance Huffman tree," IEEE Commun. Lett., vol. 24, no. 8, pp. 1602-1606, Aug. 2020.

[9] S. Halladjian, H. Miao, D. Kouril, M. E. Groller, I. Viola, and T. Isenberg, "Scale trotter: Illustrative visual travels across negative scales," IEEE Trans. Vis. Comput. Graph., vol. 26, no. 1, pp. 654-664, Jan. 2020.

[10] A. Rocha, J. D. Silva, U. R. Alim, S. Carpendale, And M. C. Sousa, "Decal-lenses: Interactive lenses on surfaces for multivariate visualization," IEEE Trans. Vis. Comput. Graph., vol. 25, no. 8, pp. 2568-2582, Aug. 2019.