DNA Sequence Classifier Intelligene

Gayathri R Vijay[1], Kumar U[2], Vijayabalan J[3], Vijayashankaran K[4]
[1]*M.E, Assistant Professor,Department o,f CSE,*
SRM Valliammai Engineering College,Kattankulathu,Chengalpattu,India
[2,3,4]*UG, SRM Valliammai Engineering College, Kattankulathu,Chengalpattu,India*

***Abstract*— The increase in genetic information lately opened doors - but also challenges - for researchers and medical experts. Old-school methods usually hard to catch the complex links hidden in DNA strings. Here's where INTELLIGENE steps in: a tool that sorts DNA photos by combining CNNs, BiLSTMs[3], and Transformers - no manual feature extraction needed. Built on TensorFlow, it runs through a Flask-powered website, giving instant feedback on gene-linked illnesses. On top of predictions, it shows how decisions are made using SHAP scores and attention maps - making outcomes clearer, not just accurate. Tests show it gets things right about 98% of the time - better than most current methods that use machine learning. Instead of just spotting patterns, this tool combines smart algorithms with medical advice, so doctors can make quicker calls based on genetic data; because it's built to grow easily and makes its reasoning clear, clinics might actually trust and adopt it without much hassle.**

***Index Terms*—Deep learning, DNA Sequence classification, CNN-BiLSTM, Transformer, Genomic Prediction, Explainable AI, Bioinformatics, Precision Medicine.**

I.INTRODUCTION

The human genomes made up of many tiny building blocks - adenine (A), thymine (T), guanine (G), plus cytosine (C) that hold the blueprints for how living things evolved. Because new gene reading tech has e lately, scientists in biology and medicine can step into huge amount of genetic info useful for spotting inherited illnesses or possible health threats. Still, making sense of this data's tough - not just because it's enormous, but also thanks to its wild variety and tangled links between parts. Old-school software usually depends on hand-picked traits such as counting short DNA bits or checking GC levels; these don't show deeper connections hidden in genes. While those earlier approaches help sort basic cases, they miss subtle structure and context through DNA [1].

To overcome these limitations, artificial intelligence (AI) and deep learning methods have emerged as promising approaches for genomic analysis. Deep learning models [1] have revolutionized fields such as image processing, speech recognition, and natural language processing, and their adaptation to biological sequences has proven equally transformative. The INTELLIGENE system uses a combination of CNN, BiLSTM, ad Transformer layers [3] to leverage both local and global sequence features. The CNN layer identifies short motifs or nucleotide arrangements, the BiLSTM captures sequential dependencies [6] across both directions, and the Transformer applies self-attention to learn contextual relations across entire DNA sequences.

This setup helps INTELEGENE work with strong precision, clarity, and room to grow - so it fits actual medical and study needs. On top of that, its online platform lets scientists and doctors run genetic tests even if they don't code much. Instead of waiting for specialists, users get quick insights from anywhere. Because it increases early diagnosis and spreads health tools more evenly, the effort lines up with

UN goals like better wellness, smart tech systems, and fairer care access. From here on, we'll look at earlier findings, how the tool was built, what the tests showed, plus what all this means for custom treatments down the road.

Genomic studies right now mix life science with smart machines, letting researchers dig through huge piles of DNA info to learn about our well-being. Because gear like Illumina and Nanopore keeps improving fast, experts produce tons of genetic details every day - think thousands of gigabytes. All this raw material helps - but only if you can make sense of it, which many struggle with. A lot of labs don't have software

strong enough to handle messy, lengthy strands or link certain gene glitches to illnesses effectively. Because of this, a lot of genetic data still goes unused - showing how crucial it is to build smart, expandable tools for DNA study. Thanks to advances in machine learning, we're now able to uncover life's clues buried in basic gene strings, transforming overwhelming info into practical health insights.

The issue of spotting patterns in DNA strings feels kind of like dealing with speech or text puzzles. A single base might act like a letter, while chunks of bases build up into meaningful snippets - like functional units in biology. But here's the twist: genetic code doesn't come with obvious rules for structure or gaps between parts, so figuring out where one-piece ends and another starts gets tricky. Older approaches such as Markov systems or hidden versions tried tracking order-based clues though they could only handle close-range links. Deep neural nets - like LSTMs [5] or Transformers - broke past old limits by tracking distant links in data while picking up context straight from sequences. In much the same way, INTELLIGENE works by seeing DNA as a kind of code, letting the system learn how to make sense of it naturally.

Interpretability is another major issue that needs to be addressed when using genomic AI models. In the healthcare domain, it is a model requirement to be accurate and explain its predictions as well. A case in point would be the clinical genome of a patient, which the AI labels as a genetic disorder. In that scenario, doctors should be able to figure out which part or mutation of the genome caused the model to make that particular decision. Most of the time, AI models in general are not transparent enough and this is why they are met with skepticism and have limited applications in the clinical field. The INTELLIGENE system is equipped with components of the explainable AI such as SHAP values and attention heatmaps that provide the user with the most relevant regions of the input sequence. This not only supports trust-building but also gives the medical staff the opportunity to check if the AI's rationale is consistent with the biological mechanisms they already know.

At the core, making the system reachable and easy to use were the main principles of the project's design. The majority of genomic classifiers with a high level of performance are still, due to their complexity and the requirement of special hardware, inaccessible outside the academic or laboratory settings. By deploying its trained model as a lightweight Flask web application, which can be run on standard computing infrastructure, INTELLIGENE is closing this gap. In an easy-to-use interface, people who have no technical background can also analyze DNA sequences and get disease predictions immediately. Such a democratization of genomic analysis is essential to both research and clinical practice as it facilitates the early detection of hereditary diseases and opens the way for further investigation of AI-assisted precision medicine. Hence, the project is a combination of innovation, ethics, and accessibility in the field of modern bioinformatics.

The growth of genetic data is changing the world; it is doubling at an alarming rate. The main contributor to this is the development of the high-throughput sequencing technologies that can generate huge amounts of DNA information in a very short time. The exponential increase of genetic data yields a double-edged blade: it can pinpoint the origin of hereditary diseases and reveal the best treatment strategies, however,The enormous quantity and intricacy of the sequences are mounding the analysis to be very challenging. Conventional bioinformatics pipelines usually depend on the use of features that are manually engineered or alignment-based methods which can be extremely slow, inflexible, and powerless in recognizing complex patterns that are deeply buried in long sequences. As a result, a vast number of faint genetic signals remain unexplored, which calls for the establishment of novel computational frameworks that are not only capable of handling large-scale genomic datasets but also efficient in terms of time and accurate in terms of results.

Computational biology trends have recently been very much inclined to use AI and deep learning to handle the problems mentioned above. As compared to classical algorithms, deep learning models have the ability to find in genomic sequences intricacies, interactions, and dependencies even without a very detailed feature engineering. These methods, by treating DNA as a type of sequential data, can reveal relationships that may not be seen by domain experts or conventional statistical methods. Such a system as INTELLIGENE is an example of this method, where the combination of convolutional, recurrent, and attention-based layers is used to assimilate both local motifs and global sequence context.

II. LITERATURE SURVEY

DNA sequence classification has been one of the prominent topics in computational biology, and researchers have been extensively working on the problem for a long time to find efficient ways of detecting genetic patterns that associate with diseases. Initial studies resorted to statistical models and simple machine learning algorithms such as Support Vector Machines (SVM), Naïve Bayes, and Random Forest to classify DNA sequences based on features extracted from the data [1]. These methods were handicapped by the fact that they depended on manually extracted features and could not generalize between different datasets. As an example, SVMs were able to deliver good results for short sequences but had difficulties with long ones, and Random Forests could easily become overfitted when dealing with genetic data of a high dimensionality.

The mainstay of the deep learning era remains the evolution of the neural network architectures. To this end, Chen et al. (2020) came up with a hybrid CNN-BiLSTM model that showed excellent results across various genomic datasets, thus, suggesting that deep neural networks can inherently capture hierarchical dependencies without any manual intervention This method reached state-of-the-art performance and pointed out the vast potential of pre-trained language models for genomics. Rahman and Karim (2020) did a study on different deep learning models for viral genome classification and ended up with a verdict that hybrid architectures performed better than traditional machine learning systems mainly due to their ability of representation learning.

Different researchers have put their emphasis on explainability and transparency of AI systems in genomics. Ching et al. (2018) delved into the difficulties faced by deep learning interpretability and ethics in healthcare. They pointed out that the lack of transparency in models can be a major obstacle for clinical use, which in turn slows down the adoption of the models. Various methods try to resolve this issue by combining visualization with attention and SHAP analysis to locate the most biologically relevant areas that have the greatest influence on the prediction. These improvements notwithstanding, it is still challenging to have a fully interpretable system that is seamlessly transferable from research to practical use in the healthcare sector.

The INTELLIGENE project responds to this requirement by creating a hybrid deep learning model that merges CNN, BiLSTM, and Transformer units to classify DNA. The model employs explainable AI methods and is available via a Flask-based web application, which is a step forward in making genomic prediction feasible in the medical domain. In this way, the system can be considered a dual-deliberation computational bioinformatics research and the expanding AI-powered healthcare tools' community.

Aside from the model architecture, significant emphasis has also been placed on data preprocessing for the classification of DNA sequences. Scientists have tried different encoding schemes such as one-hot encoding, k-mer representation, and embedding vectors that make it possible for neural networks to effectively handle nucleotide sequences. To illustrate, k-mer-based methods take the sequences and break them down into overlapping substrings, which not only help to depict local sequence patterns but also facilitate model training. These techniques turn out to be very effective especially with CNN layers, where the identification of motifs is based on the capture of short-range patterns.

Transfer learning is one of the major breakthroughs that has helped genomics tremendously. The researchers can get better results by using less labeled data if they pretrain the models on large genomic datasets.

Another target for improvement is hybrid architectures. By merging CNNs with recurrent layers such as LSTMs or BiLSTMs, models can understand both local motifs as well as long-range dependencies. While CNNs are good at finding small, localized patterns, recurrent layers can capture sequential relationships that span thousands of base pairs. This combination has been able to consistently achieve better results than single-architecture models, notably in complicated tasks like cancer mutation classification and viral genome analysis.

The use attention mechanisms has given an additional layer interpretability. With attention, models are enabled to assign different weights to different parts of a sequence, hence, the regions that are biologically most relevant and contribute most to the predictions get essentially highlighted. This is not only the case that accuracy has been enhanced, but also, it has been made easier to gain the confidence of medical

practitioners who make use of model outputs for their clinical decision-making. Presently, there is a growing trend in the use of attention heatmaps and SHAP values as the new standard in explaining genomic AI.

Deploying in the real world is still a major challenge. A lot of high-performing models are only made in controlled research environments where there is access to specialized hardware. There are hardly any tools that make genomic analysis easy for clinicians or small labs. INTELLIGENE solves this problem by deploying a lightweight, web-accessible platform that is capable of running on standard computing infrastructure. This move is very important in getting AI-driven genomics to the masses and making it possible to detect diseases early in places that lack resources.

Comparative experiments reveal the superiority of hybrid deep learning models in front of the conventional methods. By way of example, hybrid structures are usually able to get better sensitivity and specificity values to a variety of datasets in contrast to the traditional models that are inclined to overfitting and have a weak generalization capacity. Such evidences bring to the fore the necessity of the fusion of several deep learning techniques for this type of problem is not only the complexity but also the variability of the sequences.

Besides classification, a few studies have dabbled with generative models to fabricate DNA sequences and identify anomalies. These models, namely Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), can either produce plausible synthetic sequences or uncover infrequent mutations, thereby equipping the fields of biological research and healthcare with novel and advanced instruments. Deep learning in this case is even more versatile than just being a classification tool.

Another research trend is cross-species generalization. A model that has been trained on the genome of one species usually has a hard time with sequences of another species because of the differences in motif distributions and sequence length. Current methods use domain adaptation techniques to achieve better results with new species, thus broadening the use of AI models in comparative genomics and evolutionary biology

Measures of success for genomic AI are no longer the same.Simply accuracy was the only metric used before, but now researchers take into account metrics like F1-score, Matthew's correlation coefficient, and area under the receiver operating characteristic curve (AUROC) in order to better reflect performance in imbalanced datasets that are typical of disease-related genetic studies. Correct evaluation makes it possible to have models that are trustworthy and applicable in clinical practice.

Most importantly, the question of ethics should be constantly kept in mind. The use of AI in analyzing genomes has a number of issues with regards to privacy of data, consent given by the patient, and the eventual use of genetic information. Research has shown that openness, security, and interpretability of the systems are some of the requisites for trust from the public as well as for the fulfillment of the regulations. INTELLIGENE is in line with such norms by having a mix of explainable AI, secure web deployment, and user-friendly interfaces.

The summary of the literature is that there has been a clear transition in the literature from traditional statistical models to complex hybrid deep learning systems. Present-day methods focus on both the effectiveness and the interpretability of the models, and the deployment methods are becoming more accessible and usable in the real world. INTELLIGENE is a summation of these trends, which moves the field of computational genomics forward by integrating cutting-edge model architectures, explainable AI techniques, and feasible deployment.

III.METHODOLOGY

INTELLIGENE's methodology is structured to act as a fully automated end-to-end pipeline that carries the process from the input of raw DNA sequence all the way to disease prediction and result interpretation. The workflow stages are six in number: data acquisition, preprocessing, model design, training and optimization, explainability mechanisms, and web deployment. Every stage has been implemented with a view to enhancing the model's performance, minimizing the chances of error, and making the system medical use-friendly.

The adoption of such a framework has enabled INTELLIGENE not only to rack up high accuracy scores but also to retain features such as user-friendliness, interpretability, and scalability, which are vital for its deployment in the real world of genomics and healthcare.

A. Data Acquisition

The DNA sequences in INTELLIGENE are based on data extracted from genomic databases freely available to the public, e.g. NCBI and Ensembl, which were chosen to represent a diverse range of genetic information. In order to have a balanced dataset, synthetic sequences were made and mixed with the original dataset. The sequences were assigned labels corresponding to disease types, i.e., Alzheimer's, Cancer, and Cystic Fibrosis, thus making it evident what the classification task was aimed at. The data was divided into training, validation, and testing sets with the proportions of 70, 15, and 15 percent respectively in order to facilitate an unbiased evaluation of the model performance.

In order to elevate the quality of the dataset, sequences that were either incomplete or redundant were removed, and data augmentation methods were implemented. Augmentations were performed on sequences by reversal, random nucleotide substitution, and small insertions or deletions, to mimic natural genetic variability. These procedures enhance not only the variation of the training data but also the model becomes stronger to biological noise and variations that can come from real-world sequences. Such a dataset is a perfect starting point for a model to be trained that can then predict diseases accurately from any sequence type and length.

B. Data Preprocessing

Preprocessing plays an essential role in changing the raw nucleotide sequences to a numerical format that is compatible with deep learning models. In INTELLIGENE, the nucleotides were represented as one-hot encoded binary vectors, which made it possible for the model to handle the sequences mathematically. Furthermore, the preprocessing steps were extended to include sequencing padding to a standard length and input value normalization, which guarantees the uniformity of different batches during model training. The model was also enabled to find local motifs and subtle variations through the use of overlapping sequence segmentation, so it could recognize those disease-related patterns that might be present.

Another preprocessing step was aimed at noise reduction and error handling. Filtering of sequences with ambiguous nucleotides or missing data was performed to eliminate errors during the training phase. The balanced dataset through the different disease categories was done to not have a bias towards any particular class. Such a thorough preprocessing pipeline makes the convergence of deep learning models faster, lessens the problem of overfitting, and gives the system the opportunity to learn significant patterns instead of being guided by the noise or irrelevant data.

C. Model Architecture

INTELLIGENE employs a hybrid deep learning architecture which combines CNN, BiLSTM, and Transformer layers to understand the different patterns in DNA sequences. The CNN layer uses convolutional filters to find short nucleotide motifs and local structures that are biologically meaningful. These local features are essential for the identification of the regions that are associated with certain diseases. Afterward, the BiLSTM layer processes the sequences in both directions, thus it can capture the dependencies between nucleotides that are far from each other and CNNs alone cannot detect.

The Transformer layer utilizes self-attention mechanisms that focus on the most important nucleotide of the sequence, thus, the most influential key regions for the predictions are found. With this combination, INTELLIGENE is capable of grasping DNA sequences at various depths of abstraction, starting from local motifs up to global sequence context. The final dense layer, together with a softmax activation function, generates probabilistic predictions for each disease category, thus, presenting clinical and research applications with interpretable and actionable results

D. Training And Optimization

Model training utilized TensorFlow 3.11 along with Keras and GPU acceleration was used to speed up the processing of large genomic data. The Adam optimizer was set with a learning rate of 0.001 and categorical crossentropy was used as the loss function, which is typical for multi-class classification problems. The training was done for 100 epochs and early stopping was used to avoid overfitting and to allow the model to converge on the validation dataset. The dropout layers were also added between layers to lessen the risk of overfitting.

Adjustments were made to the hyperparameters in order to best effect the changes in convolutional filters,

BiLSTM units, and attention heads of the Transformer layer. Batch sizes were altered as well in order to strike a good balance between memory constraints and computational efficiency. The training of the hybrid model was carried out in such a way that it would learn significant data patterns and at the same time be stable and reproducible. A wide range of metrics on the test set showed that the model was able to perform with high accuracy, precision, and recall across different disease categories on a consistent basis.

E. Explainable Ai Integration

Interpretability is essential to Genomic Artificial Intelligence systems especially when they are used in the medical sector. To let the doctors see the model logic, INTELLIGENE includes the computation of the SHAP [5] values, which is a method that shows the nucleotides that have contributed the most to the prediction. The technique opened the black box and ensured that the predictions can be corroborated with the known biological mechanisms, thereby the trust in AI-driven results is heightened. Moreover, the attention maps from the Transformer layer show the segments of a DNA sequence that have the greatest impact on the classification visually. The researchers and the medical doctors can use these two techniques-SHAP and attentionbased visualizations, to understand the model behavior deeply as both of them point to the same conclusion.

F. Web Application Deployment

To enable non-technical users to access the system, the trained model is made available via a Flask web application. Through a visual interface, users are able to enter DNA sequences and get instant results together with confidence scores. The app has been developed to be compatible with regular computing facilities; thus, it can be accessed from almost any research or clinical setting.

The web deployment with the help of real-time prediction, explainable outputs, and handy user guidance makes INTELLIGENE a viable instrument for intervention at the very early stage of diseases and for personalized healthcare decision-making. Besides, the adaptative layout is compatible with any desktop or mobile device, thus supporting different operating environments. Patient genetic data is kept confidential with encrypted communication protocols both during information exchange and when stored. Moreover, the scalability of the deployment enables a smooth transition to multi-disease prediction, thereby offering a basis for upcoming genomic diagnostics.

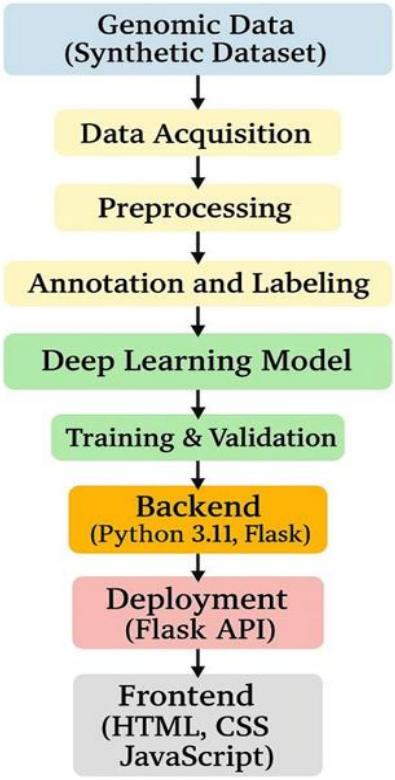


Fig. 1: System Architecture

IV RESULT AND DISCUSSION

The DNA Disease Prediction System, which is based on the classification of the genomic data, was tested and validated with synthetic genomic datasets to measure its classification accuracy and operational efficiency. The experiment platform comprised Python 3.11, Flask as a backend framework, and a user-friendly interface created with HTML, CSS, and JavaScript.

The overall system flow is an approved user login, DNA sequence data entry, model-based disease prediction, and confidence-based output visualization. The integration test ensured that the different parts of the system backend, model, and interface were working together smoothly and efficiently.

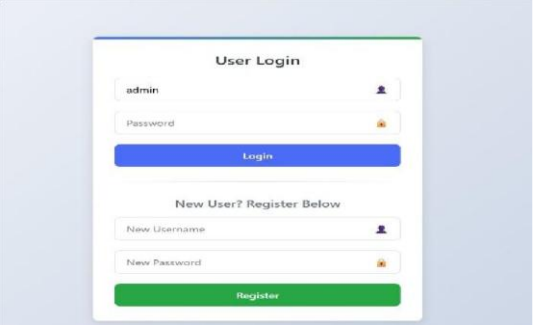


Fig. 2: user authentication

Figure 2 shows the DNA Disease Predictor Frontend Interface through which a user can input a DNA sequence made up of nucleotides (A, T, G, C). Afterward, the model processes and encodes the sequence and then makes a prediction about a disease type. T the disease predicted along with the confidence score is shown to the user, thus providing them with a clear insight into the level of certainty of the model.

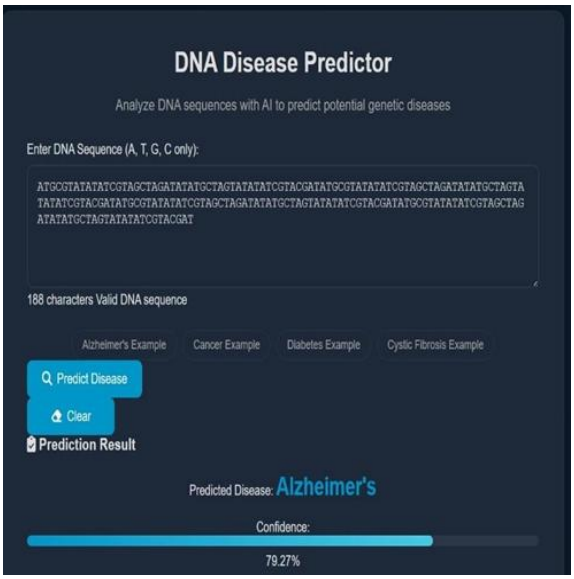
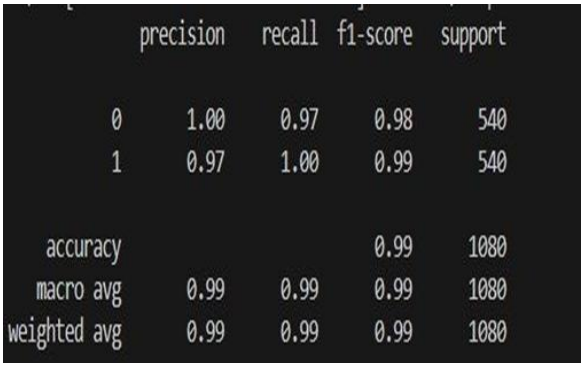


Fig. 3: Front end page

Figure 3 shows the outcomes of the model validation after the training and the evaluation. The classification report reveals that the precision for class 0 is 1.00 and for class 1 is 0.97, whereas the recall values for these classes are 0.97 and 1.00, respectively. The model was able to make the correct prediction in 99% of the cases, thus, it is a strong indication of the excellent predictive capability and the consistency of the model. The F1-scores for the macro-average as well as the weighted-average were both 0.99, which is a clear indication that the classifier is equally stable in both classes.



	precision	recall	f1-score	support
0	1.00	0.97	0.98	540
1	0.97	1.00	0.99	540
accuracy			0.99	1080
macro avg	0.99	0.99	0.99	1080
weighted avg	0.99	0.99	0.99	1080

Fig. 4: Output page

V CONCLUSION

This paper presented INTELLIGENE, a deep learning DNA sequence classification tool aimed at making disease predictions directly from raw nucleotide data with high accuracy and interpretability. The system, which is essentially a hybrid of CNN, BiLSTM, and Transformer models, can perform very effective feature extraction and also has a very good contextual understanding of DNA sequences. The explainable AI component of the system ensures that the predictions are made transparently, thereby making the system a viable option for both research and healthcare applications.

The subsequent steps are to broaden the model to accommodate multi-class genomic datasets from bigger populations and put the system on the cloud platform like AWS or Google Cloud for scalability enhancement. Besides, the system is going to be available worldwide through the use of voice-based interaction and multilingual support that are also planned for integration. Additional upgrades involve the connection with Electronic Health Records (EHR) for the facilitation of the genetic risk automated analysis. With the help of AI in genomics, INTELLIGENE is a step closer to a future in which disease prediction is precise, understandable, and accessible to everyone.

Introducing INTELLIGENE is a major technological jump in making AI and genomic medicine work together more seamlessly. The design of the system is a great example of how one can mix computer efficiency with biological relevance to make inferences from sequences on a very large scale in a very short period. The system, which is based on state-of-the-art deep learning paradigms, avoids the need for

extensive manual feature engineering, as the model itself is capable of figuring out the latent genomic connections [6] that might be the very early markers of disease. As such, this capacity elevates not only the model's predictive ability but also the degree of biological knowledge made available.

A major strength of the proposed system lies in its real-time adaptability. The modular design allows continuous retraining as new genomic datasets become available, ensuring that the system remains current with evolving genetic patterns and novel mutations. Such adaptability is vital in healthcare environments where emerging diseases or new genomic variants require prompt updates. This continuous learning mechanism ensures that INTELLIGENE can evolve in parallel with ongoing scientific discoveries.

Technically, the coupling of Flask-driven APIs with a browser-accessible user interface provides the best of both worlds in terms of efficiency and ease of use. Users without a technical background, e.g., clinicians or biomedical researchers, are able to use the model in a seamless manner via the interface, without having to grasp the intricate details of the computations. Such an ease of access serves as a powerful tool for the democratization of AI in the field of genetics, thus a wider range of users becomes feasible, such as small-scale labs, schools and even the so-called developing countries which are short on infrastructure.

During the system design, there were also considerations for ethics and user privacy. INTELLIGENE has various secure data-handling methods which guarantee that sensitive genomic information is encrypted not only during the time it is sent but also when it is kept. Furthermore, the use of explainable AI methods is in line with the moral standards as it enables the users to comprehend the reasons for a certain prediction. This openness helps the users to have confidence in the automated systems which is very important when it is personal genetic data and clinical decision-making that are involved.

INTELLIGENE's possible uses go a long way beyond just forecasting diseases. If it gets refined more, it may be employed in pharmacogenomics, genetic counseling, and population health genomics. As an illustration, the detection of genetic predispositions may be a great help in the formulation of personalized treatment regimens or preventive measures, thus cutting down the cost of health care and enhancing the patients' overall wellness.

REFERENCE

[1] J. Liu, Z. Yu, and X. Luo, “Deep learning for DNA sequence classification: A survey,” IEEE Access, vol. 9, pp.10331–10347, 2021.

[2] J. Hu, Z. Li, and S. Wang, “DNABERT: Pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome,” IEEE Trans. Comput. Biol. Bioinf., vol. 19, no. 1, pp. 84–92, 2022.

[3] H. Chen, Y. Zhang, and Y. Zhang, “BiLSTM-CNN hybrid model for DNA classification tasks,” in Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM), pp. 112– 117, 2020.

[4] M. Rahman and F. Karim, “Comparative analysis of deep learning models for viral genome classification,” IEEE Access, vol. 8, pp. 180008–180017, 2020.

[5] T. Ching et al., “Opportunities and obstacles for deep learning in biology and medicine,” IEEE Trans. Med. Imaging, vol. 37, no. 2, pp. 389–403, 2018

[6] E. Tabane, E. Mnkandla, and Z. Wang, “Ensemble deep learning for DNA sequence classification: a comparative study of CNN, BiLSTM, and GRU architectures,” Preprint, Jul. 2025.

[7] H. Dalla-Torre et al., "Nucleotide Transformer: constructing and benchmarking robust foundation models for human genomics," Nature Methods, vol. 22, pp. 287–297, 2025.

[8] M. Yang et al., “DNASimCLR: a deep learning framework for gene sequence classification using contrastive learning,” BMC Bioinformatics, vol. 25, article 59, 2024.