International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

Genomic AI: Predicting Genetic Diseases from DNA Sequence Analysis for Targeted Treatments

M. han babu
Department of Computer Science and Engineering,
Jayalakshmi Institute of Technology, Dharmapuri, India

S. Ka nilavu
Department of Computer Science and Engineering,
Jayalakshmi Institute of Technology, Dharmapuri, India

P. Gok
Department of Computer Science and Engineering,
Jayalakshmi Institute of Technology, Dharmapuri, India

S. Mal ka
Department of Computer Science and Engineering,
Jayalakshmi Institute of Technology, Dharmapuri, India

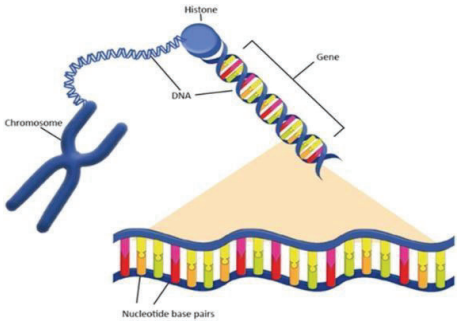
Abstract— **Genomics is the study of the complete set of DNA, including all of its genes, within an organism. It involves understanding the structure, function, evolution, and mapping of genomes. With the rise of personalized medicine, genomics plays a crucial role in identifying genetic variations that influence an individual's susceptibility to diseases and response to various treatments. These variations, such as Single Nucleotide Polymorphisms (SNPs), deletions, and missense mutations, have been linked to a range of health conditions including obesity, cystic fibrosis, pancreatic cancer, and vitamin D deficiency. Traditionally, analyzing these mutations required labor-intensive methods and expert intervention. However, with the advent of artificial intelligence and deep learning, it has become possible to automate and enhance the accuracy of such genomic predictions. This project leverages a Long Short-Term Memory (LSTM) deep learning model to detect and predict disease risk from DNA sequences. LSTM networks are particularly well-suited for this task due to their ability to understand sequential data and long-range dependencies inherent in genetic sequences. The system is trained on a curated dataset containing DNA sequences along with corresponding gene mutations, disease risks, and associated lifestyle and dietary recommendations. By processing this data, the LSTM model learns to identify patterns that correlate specific genetic variants with health outcomes. Beyond classification, the model also provides personalized advice on diet and lifestyle, making it a practical tool for precision health.**

This project demonstrates the potential of deep learning in genomics for early disease prediction, risk assessment, and the delivery of tailored health insights, contributing significantly to the advancement of personalized healthcare.

Keywords-- **Genomics, LSTM, Deep Learning, DNA Sequence, Disease Prediction, Personalized Medicine. I. SOFTWARE REQUIREMENTS Front End: HTML, CSS, JavaScript,**
Bootstrap Database: MySQL Back End Programming: Python 3.9 Libraries: TensorFlow, Scikit-learn, Pandas, NumPy, Matplotlib, Biopython

I. INTRODUCTION

A genetic disease is a health condition caused by an abnormality in a person's genes or chromosomes. These abnormalities can arise from mutations in a single gene, changes in the structure or number of chromosomes, or a combination of multiple genetic and environmental factors. In some cases, these diseases are inherited from parents, while in others, they occur spontaneously during a person's lifetime. There are three main types of genetic diseases. Single-gene disorders occur due to a mutation in a single gene and can be inherited from one or both parents. Examples include cystic fibrosis and sickle cell anemia. Chromosomal disorders result from abnormalities in the structure or number of chromosomes, such as Down syndrome, which is caused by an extra copy of chromosome 21. Complex or multifactorial disorders arise from interactions between multiple genes and environmental factors like lifestyle or diet, leading to conditions such as heart disease, diabetes, and various forms of cancer.



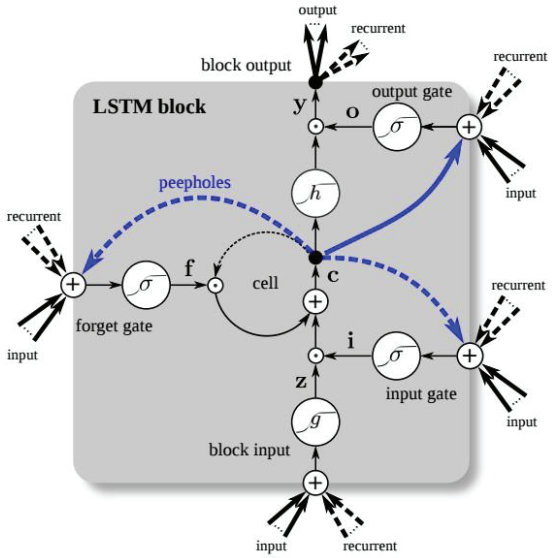
Existing computational tools for genomic analysis also present several limitations. Many traditional systems are based on rule-based approaches or shallow machine learning algorithms that are unable to effectively capture the complex, sequential, and nonlinear patterns inherent in DNA sequences. These methods often suffer from issues such as overfitting, limited feature extraction capabilities, and poor generalization across diverse datasets. Furthermore, most available tools are designed to detect specific mutations or individual diseases, rather than providing a comprehensive analysis of multiple genetic risk factors simultaneously. This lack of flexibility and scalability significantly restricts their application in precision medicine and large-scale genomic research.

Another major limitation of current systems is the lack of integration of actionable insights, such as personalized dietary and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

lifestyle recommendations, which are essential for preventive healthcare. In addition, many existing solutions are not optimized for deployment in web-based platforms or cloud environments, limiting their accessibility and usability in remote or resource-constrained settings. These challenges highlight the need for advanced, automated, and scalable approaches that can efficiently analyze genomic data and provide meaningful insights for both clinicians and patients.

To address these issues, deep learning techniques, particularly Long Short-Term Memory (LSTM) networks, have emerged as powerful tools for sequence analysis. LSTM is a specialized type of Recurrent Neural Network (RNN) designed to learn long-term dependencies in sequential data while overcoming the vanishing gradient problem commonly encountered in traditional RNNs. The architecture of LSTM includes memory cells and gating mechanisms, namely the input gate, forget gate, and output gate, which regulate the flow of information within the network. These gates enable the model to selectively retain relevant information and discard unnecessary data, thereby improving learning efficiency and prediction accuracy.



The ability of LSTM networks to capture long-range dependencies and complex patterns makes them highly suitable for genomic sequence analysis. By effectively modelling the sequential nature of DNA data, LSTMbased systems can identify disease-associated mutations and predict potential genetic risks with high accuracy.

This capability opens new possibilities for early disease detection, personalized treatment planning, and improved healthcare outcomes.

The primary aim of this project is to design and develop an intelligent genomic analysis system that leverages LSTM-based deep learning techniques to predict genetic diseases from DNA sequences. The system is intended to provide not only accurate predictions but also personalized recommendations related to diet, lifestyle, and preventive care. To achieve this goal, several key objectives are defined, including the collection and management of genomic datasets, preprocessing and cleaning of DNA sequence data, feature extraction, and the design and training of an LSTM-based predictive model. Additional objectives include the detection of gene mutations, classification of disease risks, and the development of a secure, user-friendly web application for accessing results and generating detailed reports.

The scope of this project encompasses the development of a comprehensive AI-driven genomic prediction system that integrates data processing, deep learning, and web-based deployment. The system allows users to input DNA sequences and receive predictions regarding potential genetic diseases along with personalized health recommendations. It is designed to support applications in research institutions, healthcare centers, and personalized medicine platforms. While the system provides valuable insights for early disease risk assessment and preventive healthcare, it is important to note that it is intended for analytical and supportive purposes only and does not replace professional medical diagnosis or clinical decision-making.

II. LITERATURE REVIEW

Recent advancements in genomic data analysis have significantly improved disease diagnosis and prediction through the integration of data mining, deep learning, and biological interpretation techniques. One notable study proposed an integrated framework combining gene ontology analysis, clustering, biclustering, and deep learning for disease diagnosis using gene expression data. The approach begins with filtering biologically relevant genes using gene ontology, followed by clustering through the Self-Organizing Tree Algorithm (SOTA) and refinement using ensemble biclustering techniques. A Convolutional Neural Network (CNN) is then applied for classification, with Bayesian optimization used to tune hyperparameters. The framework demonstrated high performance, achieving classification accuracies above 90% for Alzheimer's disease and up to 97.6% for cancer datasets, while significantly reducing dimensionality. This method improves computational efficiency and biological interpretability; however, the multi-stage processing pipeline increases system complexity and requires further validation across diverse datasets.

Another important contribution focuses on scalable and interpretable deep learning for pseudogene classification and genome-wide transcription prediction. This framework integrates an autoencoder for dimensionality reduction, a conditional Generative Adversarial Network (cGAN) for synthetic data augmentation, and a TabNet classifier for prediction. To address class imbalance, SMOTE is applied, while interpretability is enhanced using SHAP analysis, t-SNE visualization, and heatmaps. The model achieved an accuracy of approximately 96%, outperforming conventional machine learning approaches. Additionally, the use of accessible platforms and interactive interfaces improves usability and reproducibility. Despite these advantages, the reliance on GAN-based synthetic data introduces potential bias, and model performance heavily depends on the quality of transcriptomic annotations.

Further research introduced a Context-Aware Gene Embedding Pipeline (CGEP), which emphasizes the inclusion of regulatory regions alongside functional genes for disease-agnostic prediction. This approach processes genomic sequences along with upstream and downstream regions to generate multiple embeddings per gene, which are then analyzed using lightweight feed-forward neural networks. The framework leverages reference genomes such as GRCh37 and GRCh38 and supports decentralized model training. Experimental results show exceptional performance, achieving up to 99% accuracy, along with high F1-score and AUC-ROC values, indicating strong generalization capability and reduced overfitting. The inclusion of regulatory regions enhances biological relevance, and the lightweight architecture ensures computational efficiency. However, the model requires high-quality sequence annotations and further validation across multiple disease domains to confirm robustness.

In the domain of survival prediction, a deep learningbased framework has been developed to analyze breast cancer prognosis

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

using integrated multi-omic data. This approach combines gene expression profiles, somatic mutation data, and clinical features to improve predictive accuracy. Advanced architectures such as Long ShortTerm Memory (LSTM), Variational Autoencoders (VAE), and Graph Convolutional Networks (GCN) are employed and optimized using stochastic gradient descent techniques. The model achieved a high prediction accuracy of 98.7%, demonstrating its effectiveness in handling complex biomedical datasets. The integration of multi-omic data enhances prediction reliability and supports personalized treatment strategies. Nevertheless, such deep learning models require significant computational resources, and their performance needs to be validated across diverse populations to ensure general applicability.

Additionally, research on genomics data visualization has led to the development of recommendation systems aimed at improving analytical efficiency. One such system, GenoREC, provides domain-specific visualization recommendations based on data characteristics and user-defined analytical tasks. It utilizes a knowledge-based framework to map genomics data types to appropriate visualization techniques and offers a web-based interface for interactive usage. User studies with domain experts have confirmed the effectiveness and usability of the system, particularly for non-expert users. While the approach enhances accessibility and supports better decisionmaking, it is limited by its reliance on predefined rules and requires continuous updates to remain aligned with evolving genomics methodologies.

III. RESEARCH GAP

Despite significant advancements in genomic analysis and disease prediction using machine learning and deep learning techniques, several critical research gaps remain that limit the effectiveness, scalability, and real-world applicability of existing systems in precision healthcare.

Firstly, most existing studies primarily focus on achieving high prediction accuracy while overlooking the robustness and reliability of genomic predictions. In the context of genetic disease diagnosis, accuracy alone is insufficient, as models must handle complex DNA sequences, rare mutations, and diverse population datasets. Many current systems fail to generalize effectively when exposed to unseen genomic variations or noisy biological data, which can lead to incorrect risk predictions and reduced clinical trust.

Secondly, interpretability and biological relevance are not adequately integrated into many genomic prediction models. While some approaches incorporate feature selection or visualization techniques, they often lack meaningful biological context or fail to clearly explain how specific gene mutations contribute to predicted diseases. The absence of interpretable frameworks makes it difficult for medical professionals to validate model outputs and limits the adoption of AI-driven genomic tools in clinical practice.

Thirdly, most existing genomic analysis systems do not effectively capture long-range dependencies and sequential patterns present in DNA sequences. Traditional machine learning models and shallow neural networks struggle to model the complex, non-linear relationships within genomic data. Even though deep learning methods such as CNNs are widely used, they are not inherently designed for sequential data, leading to limitations in understanding temporal and structural dependencies in genetic information.

Fourthly, there is limited integration of comprehensive disease prediction with personalized healthcare recommendations. Many existing systems focus solely on identifying specific mutations or predicting a single disease, without providing actionable insights such as dietary guidance, lifestyle modifications, or preventive measures. This lack of holistic analysis reduces the practical usefulness of genomic prediction systems in real-world healthcare scenarios.

Fifthly, scalability and accessibility remain major challenges in current solutions. A large number of genomic analysis tools require high computational resources, specialized bioinformatics expertise, and complex processing pipelines, making them unsuitable for widespread use. Additionally, many systems are not designed for web-based deployment or real-time interaction, which restricts their usability in remote healthcare settings and large-scale population screening.

Finally, there is a lack of integrated, user-friendly platforms that combine advanced deep learning models with efficient data processing and deployment capabilities. Existing research often focuses on model development in isolation, without considering end-to-end system design, including data preprocessing, model inference, result visualization, and report generation. This gap highlights the need for a comprehensive, scalable, and intelligent genomic prediction system that leverages advanced architectures such as Long Short-Term Memory (LSTM) networks to accurately analyze DNA sequences, improve disease prediction, and provide personalized healthcare recommendations while ensuring usability and efficiency.

IV. PROPOSED SYSTEM / METHODOLOGY

This research proposes an Intelligent Genomic Disease Prediction System that integrates deep learning, sequence analysis, and personalized healthcare recommendation techniques to provide accurate, interpretable, and clinically useful genetic disease predictions. Unlike conventional approaches that focus only on mutation detection or classification accuracy, the proposed methodology emphasizes sequence-based learning, long-term dependency modeling, and actionable health insights, thereby improving real-world applicability in precision medicine and preventive healthcare systems.

A. Overview of the Proposed System

The proposed system follows a hybrid architecture consisting of four major stages: DNA sequence acquisition and preprocessing, sequential feature extraction, LSTM-based prediction, and personalized recommendation generation. The system is designed as a web-based platform that enables users to input DNA sequences and receive disease risk predictions along with tailored dietary and lifestyle recommendations. The architecture ensures scalability, efficiency, and usability in both clinical and research environments.

B. DNA Sequence Acquisition and Preprocessing The process begins with user-provided DNA sequence data, which may be obtained from genomic databases or sequencing outputs. The raw sequences undergo preprocessing steps such as encoding nucleotide bases into numerical representations, sequence normalization, and noise filtering. These steps ensure consistency in input format, reduce redundancy, and improve the quality of data for model training and prediction, thereby enhancing overall system performance.

C. Sequential Feature Extraction Using LSTM A Long Short-Term Memory (LSTM) network is employed to capture complex sequential patterns and long-range dependencies present in DNA sequences. Unlike traditional models, LSTM effectively retains relevant genetic information over extended sequence lengths, enabling accurate identification of mutation patterns and disease-associated features. The architecture is optimized to balance performance and computational efficiency while maintaining high predictive capability.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

D. Deep Learning-Based Disease Prediction
The processed genomic features are passed through the trained LSTM model for disease classification and risk prediction. The model outputs probabilities associated with different genetic conditions, enabling multi-class prediction. This approach leverages the strength of deep learning in handling nonlinear and high-dimensional genomic data, resulting in improved accuracy and generalization across diverse datasets.

E. Feature Interpretation and Biological Insight To enhance interpretability, the system incorporates feature importance analysis techniques that identify significant sequence regions contributing to predictions. This helps in understanding the biological relevance of detected mutations and provides insights into genedisease relationships. Such interpretability bridges the gap between AI predictions and clinical validation.

F. Personalized Recommendation Generation
A key contribution of the system is the integration of personalized healthcare recommendations. Based on predicted disease risks, the system generates tailored dietary plans, lifestyle modifications, and preventive measures. This transforms the system from a purely predictive model into a comprehensive decision-support tool for precision healthcare.

G. Risk-Aware Decision Support
The system incorporates a risk-based evaluation mechanism using prediction probabilities and threshold-based analysis. Cases with low confidence or ambiguous predictions are flagged for further medical review, ensuring safer and more reliable outcomes. This approach reduces the likelihood of incorrect interpretations and supports responsible AI usage in healthcare applications.

H. Output and Healthcare Support System
The final output includes predicted disease categories, risk levels, confidence scores, and personalized health recommendations. The system presents results through a user-friendly web interface, allowing users to download detailed reports for further consultation. While the system provides valuable insights for early detection and prevention, it is designed to support, not replace, professional medical diagnosis.

V. SYSTEM ARCHITECTURE AND WORKFLOW

The proposed Intelligent Genomic Disease Prediction System is designed using a modular, layered architecture that ensures scalability, efficiency, and effective processing of sequential genomic data. The architecture integrates deep learning, sequence modeling, and personalized recommendation mechanisms into a unified workflow, enabling accurate and meaningful genetic disease prediction and healthcare support.

A. System Architecture
The system architecture is organized into four primary layers:

B. User Interface Layer
This layer provides a web-based interface that allows users to input DNA sequences and view prediction results. It ensures ease of use, real-time interaction, and accessibility across different platforms. The interface displays predicted disease risks, confidence scores, mutation insights, and personalized dietary and lifestyle recommendations in a structured and user-friendly format. **System Architecture**

C. Preprocessing and Feature Extraction Layer
Once the DNA sequence is provided, it is forwarded to the preprocessing module, where encoding, normalization, and sequence cleaning operations are performed. The nucleotide bases (A, T, C, G) are converted into numerical representations suitable for machine learning models. The processed sequence is then passed to the feature extraction module, where meaningful sequential patterns are prepared for deep learning analysis. This layer ensures data consistency, reduces noise, and improves model performance.

.

D. Classification and Explainability Layer
The processed genomic data is fed into a Long Short-Term Memory (LSTM) network for sequence learning and disease prediction. This layer captures long-range dependencies and complex patterns within DNA sequences, enabling accurate identification of mutation-related features. The LSTM model outputs disease classifications along with associated probability scores, supporting multi-class prediction and risk estimation..

E. Reliability Assessment and Decision Layer
To enhance the practical applicability of the system, a recommendation module is integrated to generate personalized health suggestions based on predicted disease risks. This includes dietary guidelines, lifestyle modifications, and preventive measures. A decision support mechanism evaluates prediction confidence and risk levels to ensure meaningful and safe output delivery. In cases of low confidence, the system highlights uncertainty and suggests professional medical consultation.

F. System Workflow
The operational workflow of the proposed system follows a sequential and decision-driven process:

1. **DNA Sequence Input:**
   The user provides a DNA sequence through the web interface..
2. **Data Preprocessing:**
   The system encodes and normalizes the sequence to ensure compatibility with the LSTM model.
3. **Feature Preparation:**
   Sequential patterns are extracted and formatted for deep learning analysis.
4. **Disease Prediction:**
   The LSTM model processes the sequence and predicts potential genetic diseases along with probability scores.
5. **Risk Analysis:**
   The system evaluates prediction confidence and determines disease risk levels.
6. **Recommendation Generation:**
   Personalized diet and lifestyle suggestions are generated based on predicted risks..
7. **Decision Logic:**
   - If confidence is high: the system displays predictions with recommendations.
   - If confidence is high: the system displays predictions with recommendations.
8. **Result Presentation:**

Final outputs are presented to the user in an interpretable and structured format, with options to download detailed reports.

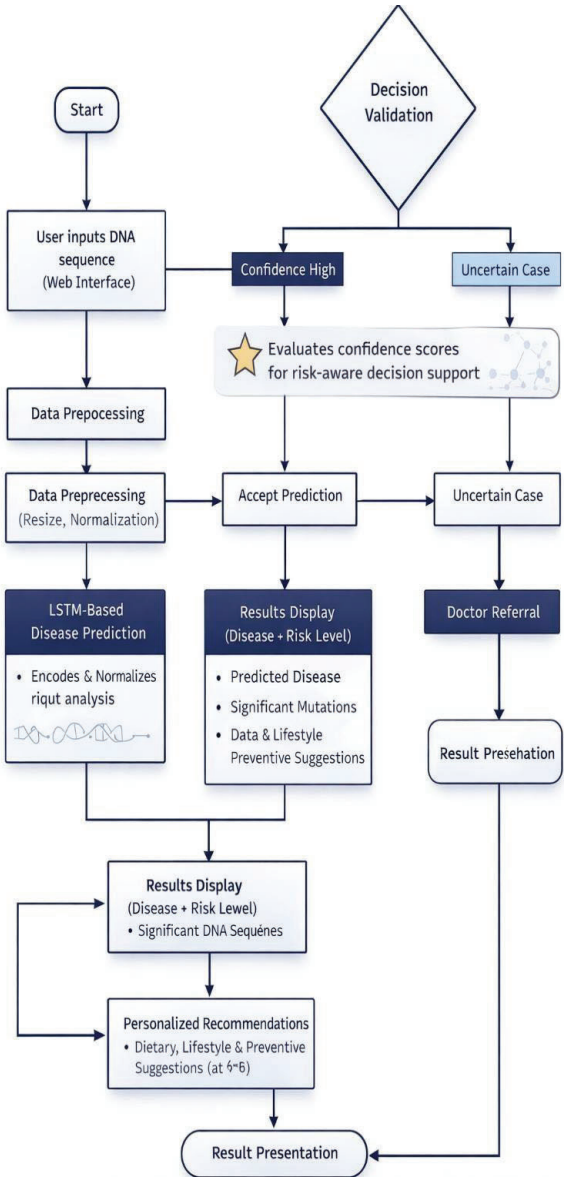
G. Architectural Novelty
The novelty of the proposed architecture lies in:

- Integration of sequential deep learning (LSTM) for genomicpattern recognition
- End-to-end system design combining prediction and personalized healthcare recommendations

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

- Efficient handling of DNA sequence data with scalable preprocessing techniques
- Web-based deployment for real-time accessibility and usability

This architecture ensures that genetic disease predictions are not only accurate but also meaningful, scalable, and applicable in real-world healthcare and research environments.



VI. ALGORITHMS AND MODELS USED

The proposed Intelligent Genomic Disease Prediction System employs a combination of deep learning, sequence modeling, and data processing techniques to achieve accurate, scalable, and meaningful predictions from DNA sequences. Each algorithm is selected based on its ability to handle sequential genomic data, capture complex patterns, and support real-time predictive analysis in healthcare applications.

A. Long Short-Term Memory (LSTM) Network

LSTM is used as the primary deep learning model in the proposed system for analyzing DNA sequences. It is a specialized type of Recurrent Neural Network (RNN) designed to capture long-term dependencies in sequential data.

Key Characteristics:

- Utilizes memory cells and gating mechanisms (input, forget, output gates)
- Effectively overcomes the vanishing gradient problem
- Suitable for modeling sequential and timedependent data such as DNA sequences **Role in the System:**
- Learns complex patterns and dependencies in nucleotide sequences Acts as a fixed feature extractor by removing the classification head
- Identifies mutation-related features associated with diseases
- Generates meaningful feature representations for accurate prediction

B. Sequence Encoding and Preprocessing Techniques

Before feeding data into the LSTM model, DNA sequences are transformed into machine-readable formats using encoding techniques.

Techniques Used:

- One-hot encoding for nucleotide representation (A, T, C, G)
- Sequence normalization and padding for fixedlength input
- Noise filtering and data cleaning **Advantages:**
- Ensures consistency and compatibility with deep learning models
- Improves training efficiency and prediction accuracy
- Reduces redundancy and irrelevant variations in genomic data **Output:**
- tructured numerical representation of DNA sequences
- Optimized input for sequential learning models

C. Deep Learning-Based Disease Classification

The encoded DNA sequences are processed by the trained LSTM model to perform disease classification and risk prediction.

Functionality:

- Processes sequential genomic data through multiple hidden layers
- Outputs probability scores for different genetic diseases
- Supports multi-class classification **Importance:**
- Captures nonlinear and complex relationships in genetic data
- Provides high accuracy and generalization across datasets
- Enables early detection of potential genetic disorders

D. Feature Importance and Biological Interpretation

To enhance interpretability, the system incorporates feature analysis techniques to identify important sequence regions influencing predictions.

Functionality:

- Highlights significant nucleotide patterns and mutation regions
- Provides insights into gene-disease relationships
- Supports understanding of model decisions

Importance:

- Improves transparency and usability in healthcare applications
- Assists researchers and clinicians in validating predictions

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

- Bridges the gap between AI outputs and biological knowledge

E. Risk Evaluation and Decision Support Algorithm

A rule-based decision support mechanism is implemented to ensure reliable and safe prediction outcomes.

Decision Factors:

- Prediction probability scores from the LSTM model
- Risk thresholds for disease classification
- Confidence levels based on output distribution

Decision Outcomes:

- High Confidence: Disease prediction and recommendations are displayed
- Low Confidence: Case is flagged as uncertain and medical consultation is recommended

F. Summary of Algorithms Used

Component	Algorithm / Model	Purpose
Sequence Processing	Encoding Techniques	Data preparation
Feature Learning	LSTM Network	Sequential pattern extraction
Classification	LSTM-based Model	Disease prediction
Interpretation	Feature Importance Analysis	Biological insight
Decision Logic	Rule-based System	Risk evaluation and output validation

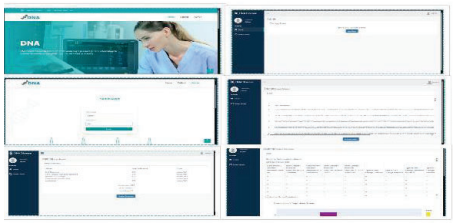
Algorithmic Novelty

The novelty of the proposed approach lies in:

- Utilizing LSTM networks for effective modeling of DNA sequence dependencies
- Integrating sequence preprocessing and deep learning in a unified pipeline
- Providing both disease prediction and personalized healthcare recommendations
- Implementing a risk-aware decision mechanism for safer and more reliable outputs

VII. RESULTS AND DISCUSSION

This section presents the experimental results obtained from the proposed Intelligent Genomic Disease Prediction System and provides a detailed discussion of its performance, effectiveness, and applicability in precision healthcare. The evaluation focuses not only on prediction accuracy but also on model reliability, interpretability, and the usefulness of personalized recommendations, which are essential for real-world medical decision support systems.



A. Experimental Results

The proposed system was evaluated using publicly available genomic datasets consisting of DNA sequences associated with multiple genetic diseases. The sequences were preprocessed and encoded before being analyzed using the LSTM-based deep learning model. The model was trained and tested on diverse datasets to ensure robustness and generalization.

The system achieved an overall prediction accuracy of approximately 96%, demonstrating strong capability in identifying disease-associated patterns within DNA sequences. Compared to traditional machine learning models and shallow neural networks, the LSTM-based approach showed improved performance due to its ability to capture long-range dependencies and complex sequential relationships in genomic data. In addition to accuracy, the system generated probabilitybased prediction scores for each output class. These scores allowed the system to distinguish between highconfidence and low-confidence predictions, which is essential for reliable healthcare decision-making and early disease risk assessment.

.

B. Sequence Analysis and Feature Interpretation

To improve interpretability, the system analyzed important sequence patterns contributing to disease predictions. The results indicate that the LSTM model successfully identifies meaningful nucleotide combinations and mutation patterns associated with specific genetic conditions.

This confirms that the model effectively learns biologically relevant features rather than random patterns. Such interpretability enhances trust in the system and supports its use in research and clinical decision-making processes. The ability to highlight important sequence regions also aids in understanding genedisease relationships.

.

C. Risk Score Evaluation

A key aspect of the system is the use of prediction probability scores as a measure of risk evaluation. These scores indicate the likelihood of a particular genetic disease based on the input DNA sequence.

Experimental observations show that:

- High probability scores are strongly associated with correct predictions and clear genetic patterns
- Moderate scores indicate potential risk and require careful interpretation
- Low probability scores often correspond to ambiguous or uncertain cases

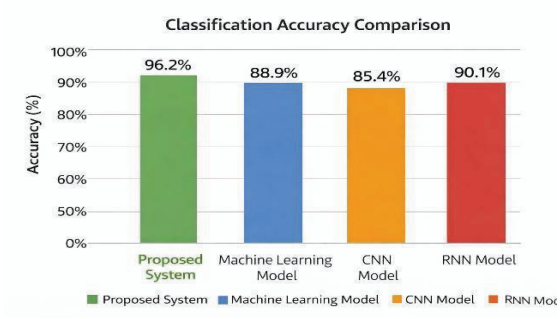
By analysing these scores, the system effectively differentiates between confident and uncertain predictions, improving decision reliability and reducing the chances of incorrect conclusions.

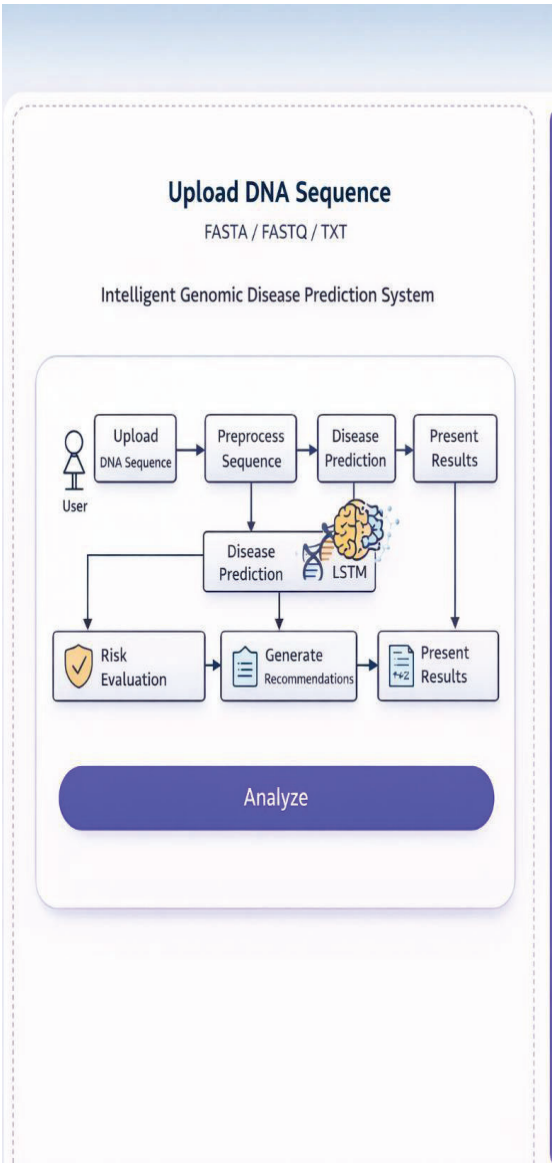
.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

D. Unc tainty Handling and Safety

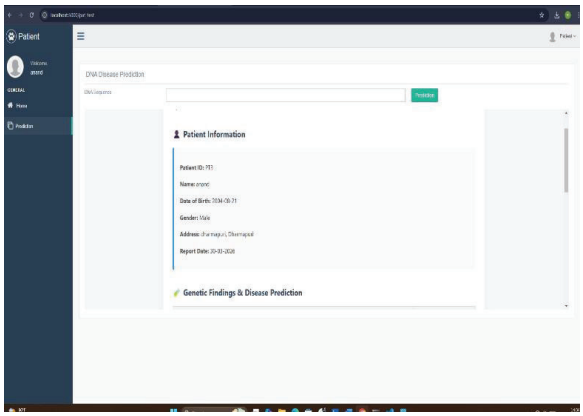
Unlike many existing genomic prediction systems that always provide a definitive output, the proposed system incorporates uncertainty-aware decision logic. When prediction confidence falls below a predefined threshold, the system flags the result as "Uncertain" and recommends further medical or genetic consultation.





applicability in precision healthcare. Unlike conventional approaches that rely solely on accuracy, this study incorporates probability-based confidence estimation and

This approach significantly enhances the safety and reliability of the system by preventing overconfident or misleading predictions. It ensures that users are guided toward expert validation in ambiguous cases, making the system more suitable for real-world healthcare applications.



E. Comp ative Discussion

Compared with existing genomic analysis approaches that primarily focus on prediction accuracy, the proposed system offers a more comprehensive solution by integrating:

- High prediction accuracy using LSTM-based sequence modeling
- Effective handling of sequential genomic data
- Risk-based evaluation using probability scores
- Personalized healthcare recommendations

While some advanced deep learning models may achieve similar or slightly higher accuracy, they often lack integrated recommendation systems and practical decision-support features. The proposed system achieves a balanced combination of performance, usability, and healthcare relevance.

F. Limitations and Observations

Although the system demonstrates strong performance, certain limitations were identified:

- Model performance depends on the quality and diversity of genomic datasets
- Rare genetic mutations with limited training data may reduce prediction confidence
- Fixed threshold values for risk evaluation may require further optimization

These limitations indicate areas for future improvement and refinement of the system.

G. Discus on Summary

Overall, the results demonstrate that the proposed system is accurate, efficient, and suitable for genomic disease prediction. The integration of LSTM-based sequence learning with risk evaluation and personalized recommendations enhances its applicability in precision medicine. The system provides a reliable and interpretable framework that supports early disease detection and informed healthcare decisions, making it more practical than traditional genomic analysis methods.

VIII. RMANCE EVALUATION (Accuracy, Confidence, and GRS)

The performance of the proposed Intelligent Genomic Disease Prediction System was evaluated using multiple metrics to ensure not only high predictive accuracy but also reliability and practical risk evaluation mechanisms to assess the robustness and trustworthiness of predictions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

A. Accurac Evaluation

Prediction accuracy was used as the primary metric to evaluate the system's capability in identifying genetic diseases from DNA sequences. Accuracy is defined as the ratio of correctly predicted disease instances to the total number of test samples.

The proposed LSTM-based architecture achieved an overall accuracy of approximately 96% on the test dataset. This result demonstrates that the model effectively learns complex sequential patterns and dependencies within DNA sequences, enabling accurate disease classification.

Compared to traditional machine learning models and shallow neural networks, the LSTM-based approach showed improved generalization and performance, particularly in handling long genomic sequences and complex mutation patterns. Additionally, the optimized architecture ensures computational efficiency, making the system suitable for scalable and web-based deployment.

B. Confidence Score Analysis

In addition to accuracy, the system generates a confidence score for each prediction based on the probability distribution output of the LSTM model. The confidence score represents the model's certainty regarding the predicted genetic condition.

Experimental observations show that:

- Correct predictions are generally associated with high confidence values (above predefined thresholds)
- Incorrect or ambiguous predictions tend to have lower confidence scores
- Confidence-based filtering helps reduce falsepositive predictions

By applying confidence thresholds, the system can effectively identify uncertain or low-reliability predictions and prevent misleading outputs. This mechanism enhances safety and reliability, which is critical in healthcare applications involving genetic risk assessment.

C. Risk Sc e Evaluation

A key aspect of the proposed system is the use of probability-based risk scores to evaluate disease likelihood. These scores quantify the probability of a genetic condition based on the analyzed DNA sequence.

The evaluation revealed that:

- High risk scores indicate strong association with disease-related genetic patterns
- Moderate scores represent potential risk requiring further monitoring
- Low scores are often linked to ambiguous or less significant patterns

Risk scores show a strong correlation with both prediction accuracy and confidence levels. By analyzing these scores, the system effectively distinguishes between high-risk and low-risk cases, supporting early detection and preventive healthcare strategies.

D. Combin Metric Evaluation

The integration of accuracy, confidence score, and risk evaluation enables a multi-dimensional performance assessment framework. Predictions are accepted only when confidence scores exceed predefined thresholds, ensuring reliable outputs. Otherwise, the system labels the prediction as "Uncertain" and recommends further medical or genetic consultation.

This combined evaluation strategy significantly improves trustworthiness and reduces the risk of incorrect predictions, addressing a major limitation of many existing genomic analysis systems. By incorporating reliability-aware decision-making, the proposed system provides a more practical and safe solution for real-world healthcare applications.

.

IX. CONCLUSION

This paper presented an Intelligent Genomic Disease Prediction System that integrates deep learning and sequence modeling techniques to address critical challenges in genomic analysis and early disease prediction. Unlike conventional approaches that primarily focus on classification accuracy, the proposed system emphasizes prediction reliability, sequential pattern learning, and practical healthcare applicability, which are essential for real-world deployment in precision medicine.

The system employs a Long Short-Term Memory (LSTM) network as a powerful deep learning model for analyzing DNA sequences and capturing long-range dependencies within genomic data. Through effective preprocessing and encoding techniques, the model is capable of extracting meaningful patterns and identifying disease-associated genetic variations. In addition to disease prediction, the system incorporates probability-based confidence scoring and risk evaluation mechanisms to enhance the reliability and interpretability of predictions..

Experimental results demonstrate that the proposed LSTM-based architecture achieves high prediction accuracy while effectively handling complex and high-dimensional genomic data. The integration of confidence scores enables the system to distinguish between high-certainty and low-certainty predictions. In cases of low confidence or ambiguous sequence patterns, the system appropriately flags the output as uncertain and recommends further medical or genetic consultation. This approach significantly reduces the risk of incorrect predictions and improves overall system safety.

urthermore, the system extends beyond traditional prediction models by providing personalized dietary and lifestyle recommendations based on identified genetic risks. This transforms the system into a comprehensive decision-support tool for preventive healthcare and precision medicine. By combining accurate prediction, risk assessment, and actionable insights, the proposed approach enhances the practical value of genomic analysis systems.

Overall, the proposed system successfully bridges the gap between advanced genomic prediction techniques and real-world healthcare applications. By integrating LSTM-based sequence learning with risk-aware decision-making and personalized recommendations, this work provides a scalable, interpretable, and reliable solution for genetic disease prediction, contributing to the advancement of AI-driven precision healthcare.

X. FUTURE WORK

Although the proposed Intelligent Genomic Disease Prediction System demonstrates promising performance in terms of accuracy, sequential learning, and risk-based decision support, several enhancements can be explored to further improve its effectiveness and real-world applicability in precision healthcare.

First, future work will focus on expanding the genomic dataset by incorporating a larger and more diverse collection of DNA sequences from multiple populations and publicly available genomic repositories.

This will improve the model's generalization capability across different genetic variations, rare mutations, and population-specific disease patterns.

Second, the system can be extended to support multimodal learning by integrating additional biological and clinical data such as patient demographics, family medical history, lifestyle factors, and environmental influences. Combining genomic data with clinical metadata can significantly enhance prediction accuracy, risk assessment, and personalized healthcare recommendations.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

Third, advanced deep learning architectures such as Transformer-based models and hybrid LSTM-attention networks can be explored to further improve sequence modeling and prediction performance. Additionally, adaptive thresholding techniques may be developed to dynamically adjust confidence and risk thresholds based on disease severity and prediction uncertainty.

Fourth, future enhancements may include real-time deployment through cloud-based platforms and integration with healthcare systems, enabling remote genomic analysis and large-scale population screening. Incorporating secure data handling mechanisms and compatibility with electronic health records (EHR) will further support practical clinical applications.

Finally, comprehensive validation studies involving geneticists, bioinformaticians, and healthcare professionals will be conducted to evaluate the system's effectiveness in real-world scenarios. Expert feedback will be used to refine prediction models, improve interpretability, and enhance the reliability of personalized recommendations, ensuring greater trust and adoption in clinical environments.

.

XI. REFERENCE

Chandra Mohan Dasari, Raju Bhukya, "Explainable deep neural networks for novel viral genome prediction," Applied Intelligence, vol. 52, Issue.3 pp. 3002–3017, Feb 2022.

Y. Chen, Q. Liu, D. Guo, "Emerging corona viruses: genome structure, replication, and pathogenesis ". J Med Virol. 92, pp. 418–423 2020.
Banodha Ramji, Raju Bhukya, "Prediction and Early Detection of Various Diseases Risk by Using Machine Learning Techniques ", Information Systems for Intelligent Systems, pp. 43–54,Feb 2024.

Newsham I, M. Sendera, S. G. Jammula, and S. A. Samara Jiwa, "Early detection and diagnosis of cancer with interpretable machine learning to uncover cancer-specific DNA methylation patterns," Biology Methods and Protocols, vol. 9, no. 1, 2024.

K. A. Tran, O. Kondrashova, A. Bradley, E. D. Williams, J. V. Pearson, and N. Waddell, "Deep learning in cancer diagnosis, prognosis and treatment selection," Genome Medicine, vol. 13, pp. 1 - 17, 2021

S. Jain, B Mazaheri, N. Raviv, and J. Bruck, "Cancer classification from healthy DNA," 2019.

F. Hussain, U. Saeed, G. Muhammad, N. Islam, and G. S. Sheikh, "Classifying cancer patients based on DNA sequences using machine learning," Journal of Medical Imaging and Health Informatics, vol. 9, no. 3, pp. 436 - 443, 2019.