IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.1412108

NextGenAI Genomic Biomarker System: A Hybrid Machine Learning Approach for Early Genetic Disorder Detection

Bhavana Suresh[1], Greeshma R Gowda[2], Dr.Abhilash C N[3]

Student, AIML, SJB Institute of Technology, Bengaluru, India[1]

Student, AIML, SJB Institute of Technology, Bengaluru, India[2]

Professor and HOD, AIML, SJB Institute of Technology, Bengaluru, India[3]

Abstract: The interpretation of vast genomic datasets remains challenging due to complexity and cognitive burden on clinicians. The NextGen AI Genomic Biomarker System addresses these challenges through a hybrid architecture combining NLP and Deep Learning. The system leverages TF-IDF vectorization with Random Forest classification achieving weighted F1-score of 0.874, and employs CNN-LSTM architecture achieving AUC of 0.93. Integrated with SHAP-based explainability, the system provides transparent predictions with sub-2-second latency while maintaining HIPAA/GDPR compliance. Index Terms—Genomic biomarkers, precision medicine, NLP, deep learning, explainable AI, Random Forest, CNN-LSTM.

I. INTRODUCTION

A. Overview

The 21st century has witnessed transformative advances in genomics and personalized medicine. As sequencing technologies become exponentially faster and cheaper, healthcare is shifting from reactive, symptom-driven diagnostics toward proactive, prediction-oriented models where genetic information guides clinical decisions before disease manifestation. This transition marks the beginning of genomic intelligence: the integration of biological datasets with advanced artificial intelligence to decode the genetic foundations of disease. However, genetic reports produced by nextgeneration sequencing (NGS), whole genome sequencing (WGS), or variant analysis pipelines are rich with data but poor in interpretability. They contain complex scientific language, biomarker descriptions, variant categorizations, and layered interpretations that require significant domain expertise to evaluate. As a result, clinicians, genetic counselors, and researchers regularly encounter barriers in converting these reports into actionable insights, creating bottlenecks in early disease detection, intervention planning, and personalized treatment pathways. The NextGen AI Genomic Biomarker System is conceptualized as an innovative response to these modern diagnostic challenges. Positioned at the intersection of computational genomics and natural language processing, the system leverages machine learning to analyze unstructured genomic reports, extract clinically meaningful biomarkers, and predict early stage genetic disorders.

B. Background of the Study

The current standard of care suffers from three major interconnected crises:

- *1) The Interpretation Gap:* Clinicians and geneticists are overwhelmed by the sheer number of variants identified in a single genome, often millions of single nucleotide polymorphisms (SNPs). Distinguishing between a benign polymorphism and a truly pathogenic variant of unknown significance (VUS) is a complex, time-intensive task. A significant portion of novel variants fall into the VUS category, requiring manual literature review to assess pathogenicity. This manual curation is slow, subjective, and creates a significant backlog, limiting the diagnostic yield of sequencing.
- *2) Multi-Modal Data Fragmentation:* Diagnostic information in modern hospitals is highly fragmented across structured genomic data (VCF, BAM) and unstructured clinical records such as EHR notes, patient histories, and pathology reports. These isolated data silos prevent a holistic diagnostic view, as current bioinformatics tools focus on sequence analysis while ignoring rich contextual information in textual records.
- *3) The Diagnostic Odyssey:* The cumulative effect of the interpretation gap and data silos is the protracted, often years long process a patient with a rare or complex genetic disease endures before receiving a definitive diagnosis. This delays effective treatment, increases patient suffering, generates immense psychological stress for families, and wastes healthcare resources on non-specific, ineffective treatments.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.1412108

C. Problem Formulation

The project establishes specific, quantifiable technical and functional objectives:

1) Technical Objectives:

- Engineer a robust TF-IDF/Random Forest pipeline for text-based disorder prediction (Target: Weighted F1Score ≥ 0.85).
- Integrate a CNN-LSTM hybrid architecture for variant pathogenicity prediction (Target: AUC ≥ 0.90).
- Embed SHAP framework for explainable predictions with visual explanations. Achieve prediction latency < 2.0 seconds for 95th percentile of requests.

2) Functional Objectives:

- Develop multi-modal ingestion module processing PDF/DOCX and VCF/BAM files.
- Implement PII masking and data sanitization for regulatory compliance. Ensure HIPAA/GDPR compliance with robust security architecture.

D. Significance of the Study

This study addresses the critical gap between genomic data and clinical context by integrating AI-driven analysis of both structured genomic files and unstructured medical reports. It enables early, interpretable prediction of genetic disorders, supporting faster diagnosis and improved clinical decision-making.

E. Scope and Limitations

The project focuses on early-stage genetic disorder prediction using NLP and machine learning on selected genomic and clinical datasets. Its performance is limited by data quality, dataset diversity, and availability of well-labeled reports,and it is intended as a decision-support tool rather than a definitive diagnostic system.

II. LITERATURE REVIEW

A. Evolution of Genomic Analysis Methodologies

The methods for identifying disease-causing genes have rapidly evolved over the past two decades, moving from simple statistical correlation to complex deep neural network modeling.

1) Early Statistical Methods and GWAS: The initial breakthroughs in complex disease association were largely driven Genome-Wide Association Studies (GWAS). GWAS rapidly scans markers across the entire genome of many individuals to find genetic variations associated with particular diseases, typically employing simple linear statistical models such as logistic regression to test the association between each individual Single Nucleotide Polymorphism (SNP) and the phenotype. However, GWAS is effective only for identifying common variants with modest effects in large cohorts. It fails when dealing with two critical issues that contribute to the "missing heritability" problem: rare variants that are not common enough to pass stringent statistical significance thresholds, and epistasis (gene-gene interactions) where the core assumption of additive variant effects breaks down entirely.

1) The Rise of Deep Learning in Genomics: The inherent limitations of GWAS, particularly in handling the high dimensionality of millions of variants and non-linearity of genomic data, necessitated a shift toward machine learning. The research by Poplin et al. (2018) marked a pivotal moment, demonstrating that specialized Deep Learning architectures could achieve expert-level accuracy in variant calling from raw sequence data. The innovation lay in treating aligned genomic data (sequence reads mapped to the reference genome) as 2D images, allowing Convolutional Neural Networks (CNNs) to learn patterns of sequence variation directly from raw data, bypassing the need for manual feature engineering.

B. NLP in Clinical Genomics

A major inefficiency in genomic diagnostics is the manual effort required to link a patient's genetic variants to their clinical symptoms. Phenotypic data (symptoms, severity, age of onset), often contained in unstructured clinical notes, is crucial for prioritizing the analysis of millions of genetic variants. Wang et al. (2021) demonstrated that integrating textual EHR data with sequencing information significantly improved diagnostic yield, highlighting the value of NLP in genomics.

While modern NLP is dominated by large, pre-trained transformer models such as BERT and GPT-4, this project adopts a more efficient and clinically transparent methodology using TF-IDF Vectorization and Random Forest Classification for three strategic reasons:

1) *Computational Efficiency:* Transformer models require massive GPU resources and are prone to high inference latency, making them challenging for real-time clinical deployment. TF-IDF and RF are highly optimized, CPU-friendly algorithms designed to achieve the target latency of 2.0 seconds.

2) *Interpretability Advantage:* The Random Forest model offers higher inherent transparency than Deep Neural Networks. Its decision-making logic is directly traceable: TF-IDF explicitly measures the discriminative power of

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.1412108

specific tokens by weighting them based on their frequency in a document and rarity across the corpus, and RF uses these term weights directly to make decisions.

3) N-gram Modeling: The use of N-grams (up to 3-grams) within the TF-IDF scheme is crucial. Many biomarkers are multi-word phrases (e.g., "exon 5 deletion", "cystic fibrosis transmembrane regulator") that must be treated as single, continuous features to maintain semantic integrity.

C. Deep Learning for Sequence Analysis

For high-resolution DNA sequence analysis, the system employs a hybrid deep learning architecture capable of capturing both local and long-range genomic patterns. Convolutional Neural Networks (CNNs) effectively detect biologically significant motifs such as promoters and splice sites from one-hot encoded sequences. Long Short-Term Memory (LSTM) networks model long-range regulatory dependencies across the genome, enabling contextual understanding of gene regulation. The CNN–LSTM combination allows integrated learning of local motifs and their broader regulatory interactions.

D. Explainable AI: The Clinical Imperative

The use of AI in medical diagnostics requires transparency and accountability to avoid the "black box" problem, where predictions lack clear justification. Clinicians must be able to understand and audit model decisions to ensure ethical and legal reliability. To address this, the system employs SHAP, which explains individual predictions by quantifying the contribution of each feature. This provides clear, patient-specific insights that support trustworthy and actionable clinical decision-making.

III. SYSTEM DESIGN AND ANALYSIS

A. Architecture

Layered microservices with independent scalability: **Presentation** (React.js, JWT, MFA); **Application** (Flask API with /auth, /predict, /reports); **Intelligence** (NLP/RF Service-CPU, DL/CNN-LSTM Service-GPU, XAI Service); **Data Persistence** (PostgreSQL, S3 object storage).

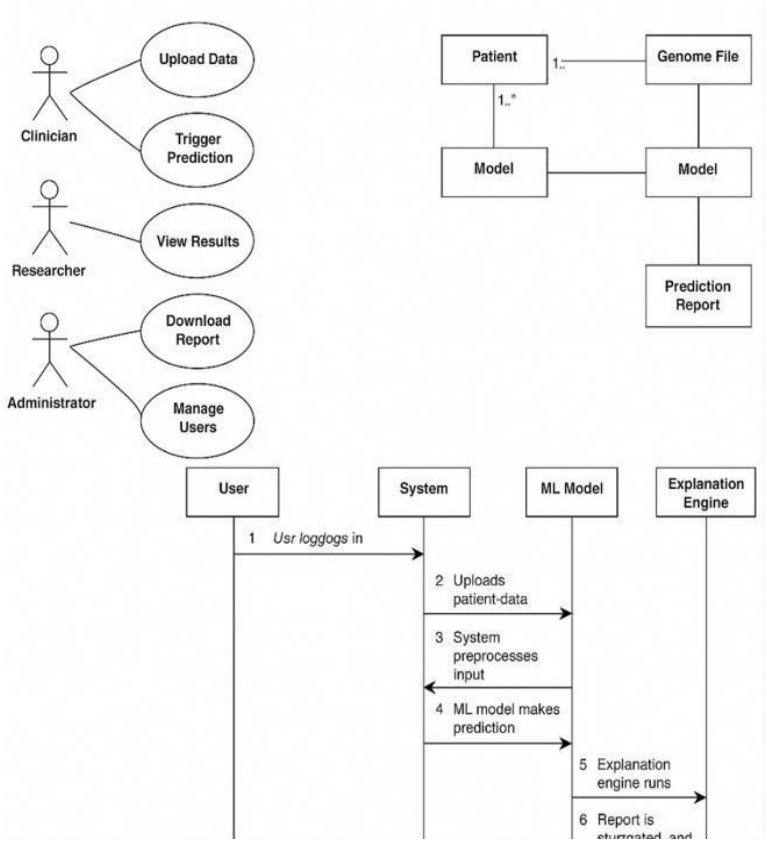


Fig. 1. Use Case Diagram

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.1412108

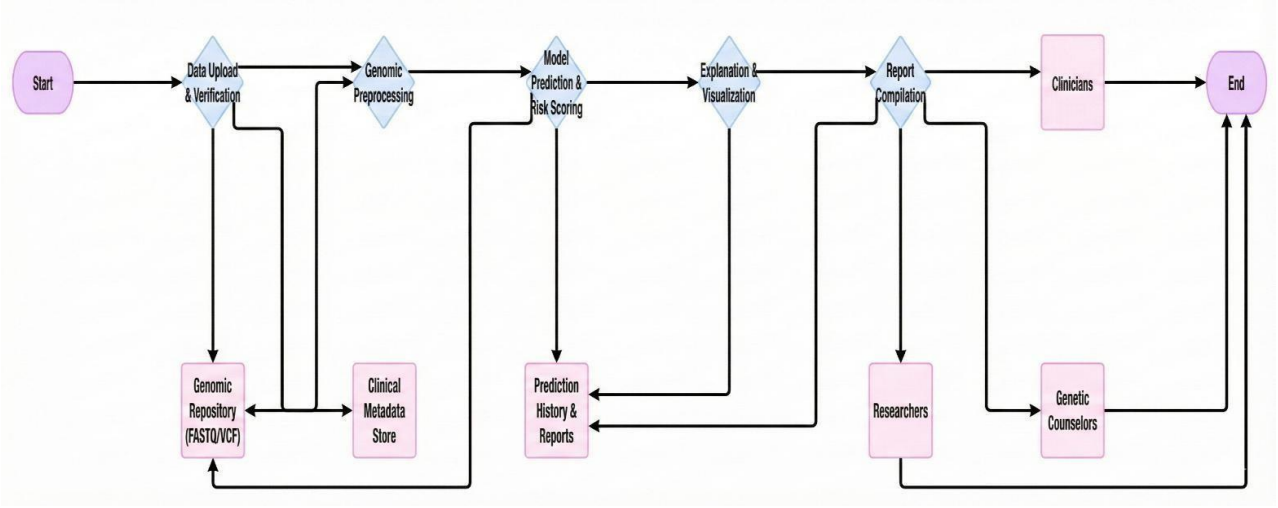


Fig. 2. Data Flow Diagram

B. Requirements

- Hardware: Intel Xeon Scalable (16+ vCPUs), 128GB RAM, 20TB NVMe SSD + S3, NVIDIAA100 GPU.
- Software: Python 3.10+, Flask, PostgreSQL 14+, Celery/Redis, Docker, Kubernetes. Security: AES-256-GCM (at rest), TLS 1.3 (in transit), three-tier RBAC, PII masking, HIPAA/GDPR compliance.

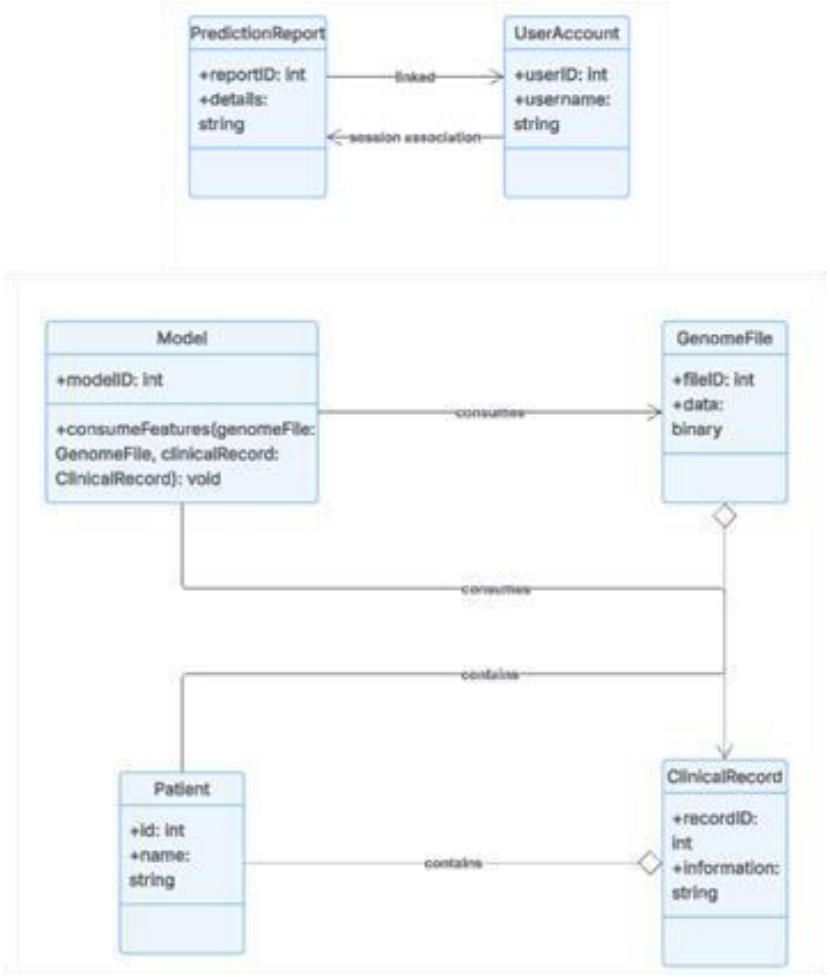


Fig. 3. Class Diagram

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.1412108

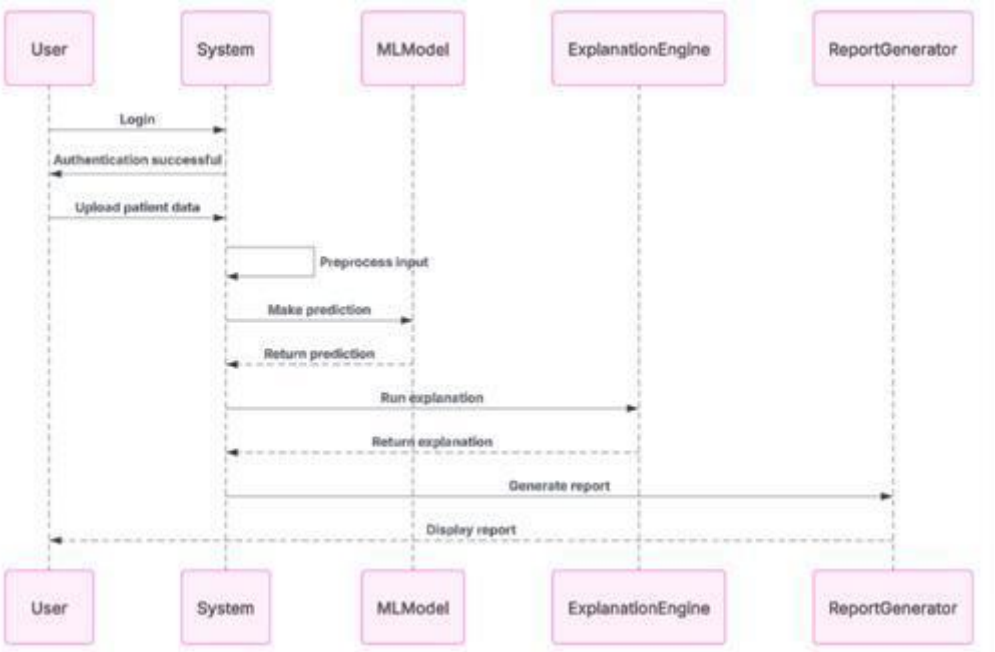


Fig. 4. Sequence Diagram

IV. MACHINE LEARNING METHODOLOGY

A. Text-Based Phenotype Analysis

TF-IDF vectorization transforms 5,000+ reports into sparse matrix (99.8% zeros) with 3,000 features using N-gram range (1,3). IDF heavily penalizes common terms, prioritizing rare biomarkers. Trigrams preserve multi-word entities. Random Forest uses parallel ensemble of decorrelated trees via bagging, random feature subsets, Gini impurity splits. Hyperparameters: n estimators=150, max depth=20, class weight='balanced subsample' for imbalance handling.

B. Sequence-Based Genotype Analysis

VCF inputs transformed to fixed-length one-hot encoded tensors.

CNN: Convolutional filters (8-16 bases) detect motifs (TFBS, splice sites); max-pooling reduces dimensionality.

LSTM: Processes CNN features modeling long-range dependencies via gated mechanisms.

Output: Dense layer with SoftMax generates probabilities (Pathogenic, Benign, VUS) using Categorical Cross-Entropy loss.

C. Feature Fusion

F_text (RF probabilities) and F_seq (CNN-LSTM outputs) concatenate to F_combined. Final dense layers learn nonlinear phenotype-genotype relationships.

V. IMPLEMENTATION AND TESTING

Libraries: NLTK/spaCy (preprocessing), Pandas/NumPy (data), Pickle (serialization), TensorFlow/Keras (DL), Scikitlearn (RF), SHAP (XAI).

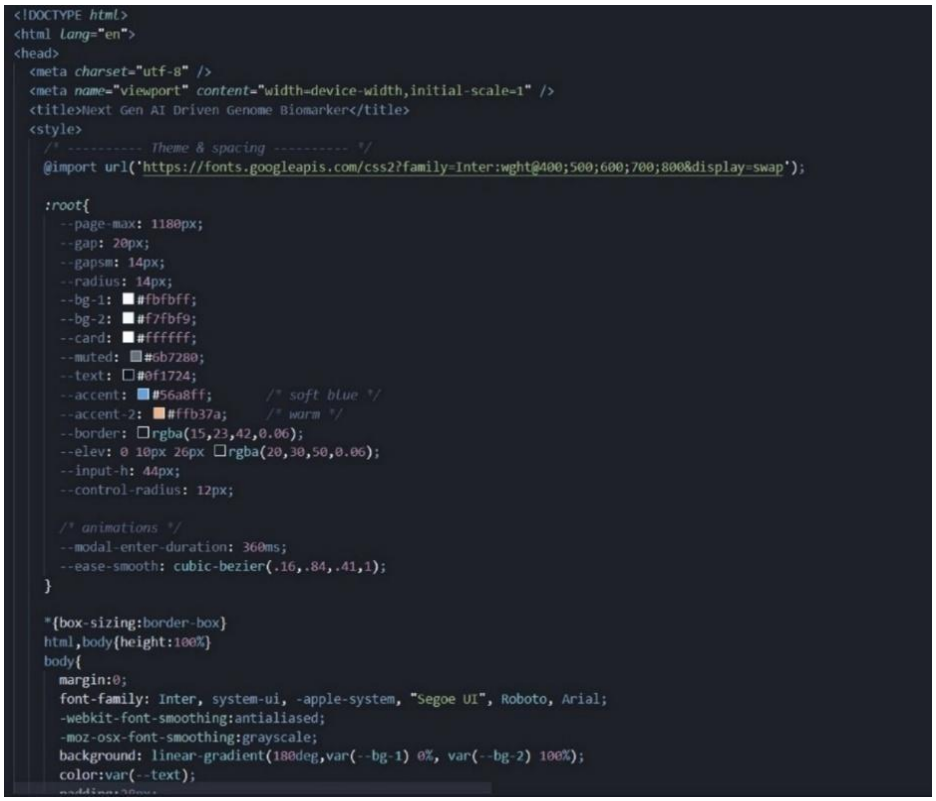
Deployment: Docker containers with K8s orchestration, HPA on CPU, GPU scheduling, 99.9% uptime, Celery workers for async tasks.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.1412108



```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="utf-8" />
  <meta name="viewport" content="width=device-width,initial-scale=1" />
  <title>Next Gen AI Driven Genome Biomarker</title>
  <style>
    /* ---------- Theme & spacing ---------- */
    @import url('https://fonts.googleapis.com/css2?family=Inter:wght@400;500;600;700;800&display=swap');

    :root{
      --page-max: 1180px;
      --gap: 20px;
      --gapsm: 14px;
      --radius: 14px;
      --bg-1: #fbfbff;
      --bg-2: #f7fbf9;
      --card: #ffffff;
      --muted: #6b7280;
      --text: #0f1724;
      --accent: #56a8ff;        /* soft blue */
      --accent-2: #ffb37a;      /* warm */
      --border: rgba(15,23,42,0.06);
      --elev: 0 10px 26px rgba(20,30,50,0.06);
      --input-h: 44px;
      --control-radius: 12px;

      /* animations */
      --modal-enter-duration: 360ms;
      --ease-smooth: cubic-bezier(.16,.84,.41,1);
    }

    *{box-sizing:border-box}
    html,body{height:100%}
    body{
      margin:0;
      font-family: Inter, system-ui, -apple-system, "Segoe UI", Roboto, Arial;
      -webkit-font-smoothing:antialiased;
      -moz-osx-font-smoothing:grayscale;
      background: linear-gradient(180deg,var(--bg-1) 0%, var(--bg-2) 100%);
      color:var(--text);
```

Fig. 5. Front End Implementation

```
from sklearn.pipeline import Pipeline
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.ensemble import RandomForestClassifier
import joblib
import pandas as pd

def train_and_save_model():
    df = pd.read_csv('reports_dataset.csv')
    X = df['report_text']
    y = df['disorder']

    pipeline = Pipeline([
        ('tfidf', TfidfVectorizer(
            stop_words='english',
            lowercase=True,
            ngram_range=(1,3),
            max_features=5000,
            max_df=0.8,
            min_df=2
        )),
        ('clf', RandomForestClassifier(
            n_estimators=200,
            max_depth=15,
            random_state=42,
            class_weight='balanced'
        ))
    ])

    print("Training model...")
    pipeline.fit(X, y)

    print("Saving trained model...")
    joblib.dump(pipeline, 'genetic_disorder_model.pkl')
    print("✅ Model retrained and saved successfully!")

if __name__ == "__main__":
    train_and_save_model()

import joblib

class TextDisorderPredictor:
```

Fig. 6. ML Model Implementation

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.1412108

```python
@app.route('/api/analyze', methods=['POST'])
def analyze():
    try:
        if model is None:
            return jsonify({'error': 'Model not available on server.'}), 500

        if 'report' not in request.files:
            return jsonify({'error': 'No report file provided.'}), 400

        file = request.files['report']
        if not file.filename.lower().endswith('.pdf'):
            return jsonify({'error': 'Invalid file type. Please upload a PDF.'}), 400

        # Extract text from PDF
        pdf_stream = io.BytesIO(file.read())
        reader = PdfReader(pdf_stream)
        extracted_text = ""
        for page in reader.pages:
            extracted_text += page.extract_text() or ""

        if not extracted_text.strip():
            return jsonify({'error': 'Could not extract readable text from the PDF.'}), 400

        # Model inference
        prediction = model.predict([extracted_text])[0]
        # predict_proba may return probabilities; ensure we get the max *percentage*
        try:
            confidence = float(max(model.predict_proba([extracted_text])[0]) * 100)
        except Exception:
            # fallback if predict_proba not available
            confidence = 0.0

        details = DISORDER_DETAILS.get(prediction, DISORDER_DETAILS['default'])

        # Save result to database (non-blocking for API flow)
        filename = getattr(file, 'filename', 'uploaded_report.pdf')
        requested_by = request.form.get('requested_by') or None
        requested_role = request.form.get('requested_role') or None
        # Save inside try/except via helper (which logs errors)
        save_prediction_record(filename, prediction, confidence, requested_by, requested_role, extra=None)
```

Fig. 7. API Endpoint

```python
import io
import os
import time
import traceback
import sqlite3
from datetime import datetime
from flask import Flask, request, jsonify, send_from_directory, g
from flask_cors import CORS
from PyPDF2 import PdfReader
import joblib

# --- App Initialization ---
app = Flask(__name__, static_folder='static')
CORS(app)

# --- Model Load ---
print("-" * 60)
print("Initializing AI Model...")

MODEL_PATH = "genetic_disorder_model.pkl"
try:
    model = joblib.load(MODEL_PATH)
    print(f"✅ Model loaded successfully from {MODEL_PATH}")
except Exception as e:
    print(f"❌ Failed to load model: {e}")
    model = None

print("🚀 Real Text-Based Analysis Server Started")
print("Backend is running at: http://localhost:5000")
print("-" * 60)

# --- Disorder Explanations & Management ---
DISORDER_DETAILS = {
    'Healthy': {
        'explanation': (
            "The text analysis indicates no significant markers for the genetic "
            "disorders in our panel. The report's language aligns with that of a healthy "
            "individual, suggesting a low probability of these specific conditions."
        ),
        'management': (
            "Maintain a healthy lifestyle with regular check-ups, balanced diet, and "
```

Fig. 8. Backend Implementation

A. Methodology

85% code coverage using pytest. **Unit:** PII masking (100% replacement), serialization integrity (identical predictions). **Integration:** Full workflow (Auth → Ingestion → Masking → DB → ML → Audit), async handling (HTTP 202), RBAC enforcement.

Validation: 5-Fold Stratified CV, Weighted F1 (primary), MCC, AUC. SHAP tested with synthetic features.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.1412108

System: Locust (250 users, 1.6s at 95th percentile), GPU cluster (20 WES/hour), stress (graceful degradation), security (penetration testing).

TABLE I: GDPR/HIPAATechnical Controls

Reg.	Mandate	Control
HIPAA	ePHI Prot.	RBAC, MFA, Audit Logging
GDPR 17	Erasure	Cascade deletion scripts
Both	Encryption	TLS 1.3, AES-256-GCM
GDPR 25	Privacy	PII Masking (NER)

TABLE II: Test Cases

ID	Module	Description	Expected	Status
1	UI	Login	Entry form	Pass
2	Text	Convert	Numeric	Pass
3	Analysis	Input	Analysed	Pass
4	Session	Close	Closed	Pass
5	Validate	Map	Match	Pass
6	Predict	Model	Output	Pass

VI. RESULTS AND ANALYSIS

A. Text Pipeline Performance

TABLE III: Random Forest Performance

Metric	Result	Target	Status
Weighted F1	0.874	$\geq$0.85	Met
Recall	0.891	Max	High
Precision	0.870	N/A	High
Latency (95%)	1.6s	$\leq$2.0s	Met

Weighted F1-score of 0.874 confirms balanced performance across 50 disease classes. High Recall (0.891) critical for clinical safety, minimizing false negatives.

B. SHAP Explainability

Global: Top features were N-grams (“cystic fibrosis transmembrane regulator”, “exon skipping mutation”), validating strategy. Non-specific terms near zero.

Local (DMD): Positive drivers – “dystrophin gene” (+0.42), “frame shift deletion” (+0.38); Negative – “female patient” (-0.15); Final 98% confidence from cumulative +1.09 overwhelming -0.15, providing auditable reasoning.

C. Sequence Pipeline

AUC 0.93 with 91.5% sensitivity for pathogenic variants. CNN detected local motifs, LSTM understood long-range effects, addressing VUS challenge.

D. System Implementation

The proposed system provides a secure and intuitive clinical analytics dashboard that enables seamless navigation across genomic data ingestion, AI-driven prediction, explainable insights, and diagnostic report generation modules.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.1412108

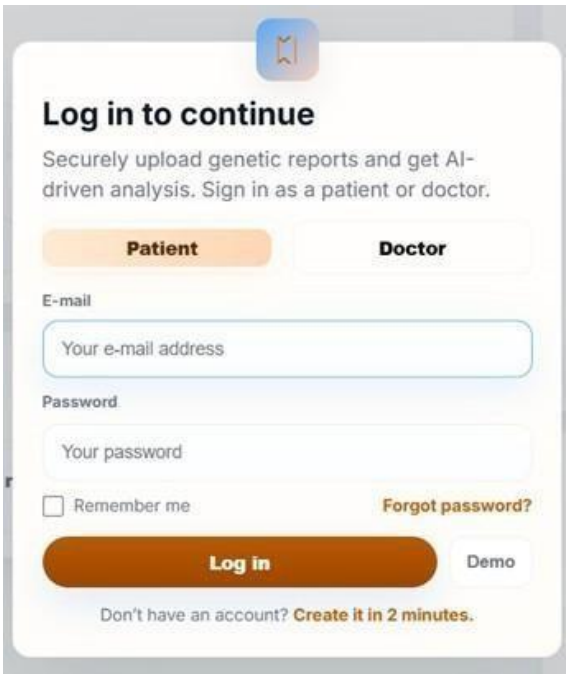


Fig. 10. Login Page

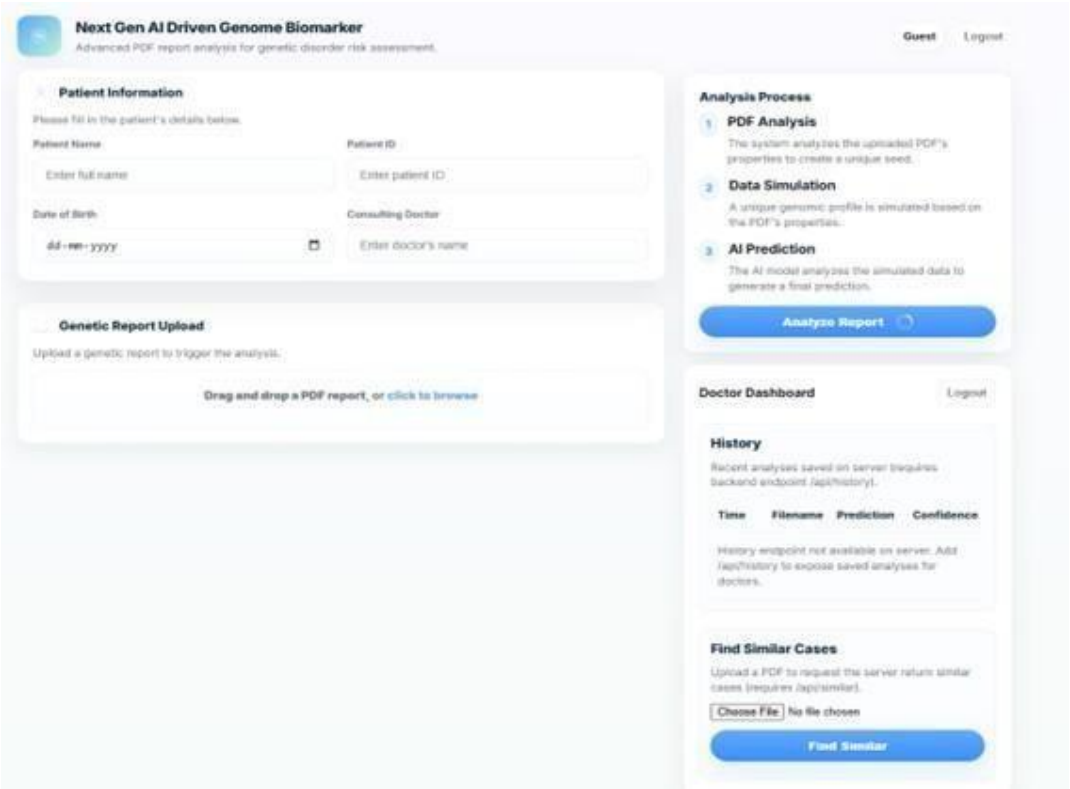


Fig. 10 Home Page - Main Interface

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025

DOI: 10.17148/IJARCCE.2025.1412108

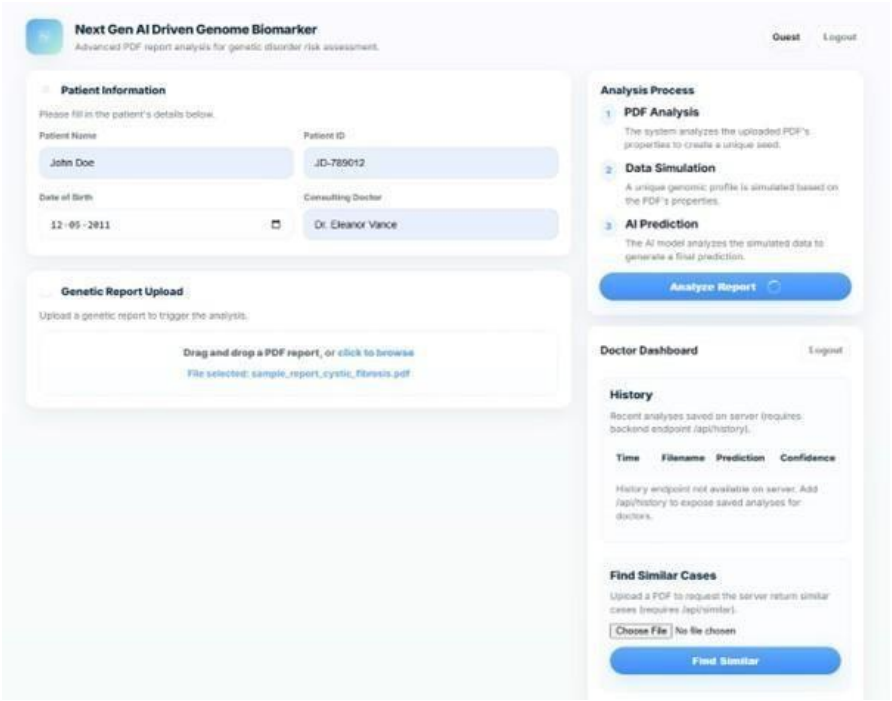


Fig. 11. Dashboard Features

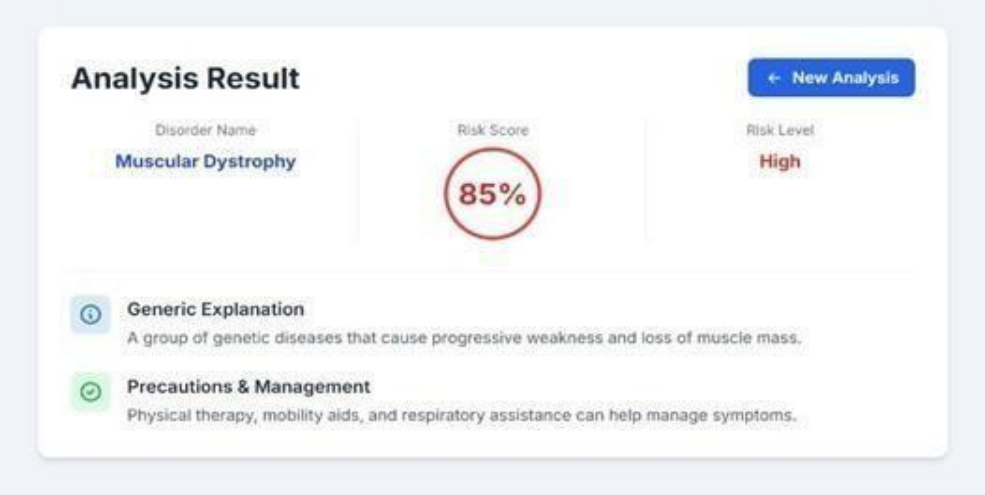


Fig. 12. Analysis Interface - Report Processing

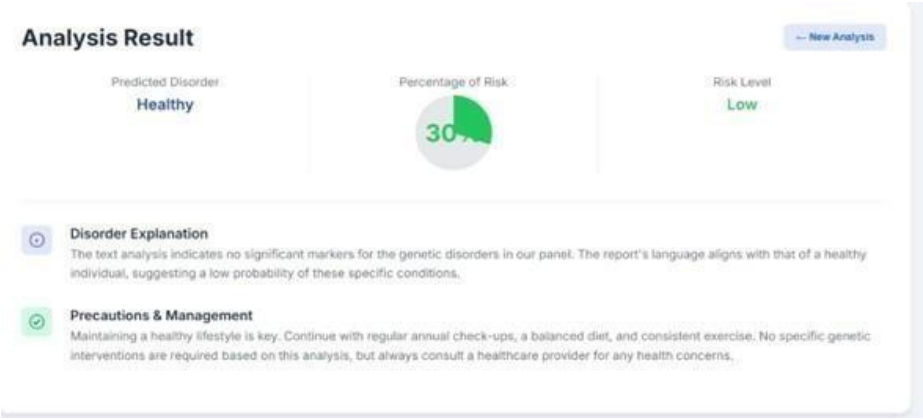


Fig. 13. Analysis Results with SHAP Visualization

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 12, December 2025
DOI: 10.17148/IJARCCE.2025.1412108

VII. CONCLUSION AND FUTURE WORK

A. Achievements

Successfully delivered secure, high-performance, interpretable AI platform addressing genomic interpretation gap. Key results: Weighted F1 0.874 (exceeding 0.85 target) with Recall 0.891 for clinical safety; AUC 0.93 with 91.5% sensitivity; 1.6s latency (below 2.0s); SHAP explainability with auditable reasoning; HIPAA/GDPR compliance (AES256- GCM, TLS 1.3, RBAC); successful imbalance mitigation via stratified CV and balanced weighting.

B. Clinical Impact

Accelerates diagnosis by automating phenotypic-genotypic interpretation as Clinical Decision Support System. Flags high risk cases missed due to cognitive load. Enables early detection and preventive interventions. Builds clinical trust through transparent reasoning. Reduces expert hour costs, minimizes iterative testing, optimizes resource allocation.

C. Scientific Contribution

Validates novel phenotype-genotype fusion methodology demonstrating high-performance without expensive LLMs. Modular framework proves efficient, interpretable architectures (TF-IDF/RF for text, CNN-LSTM for sequences) achieve clinical-grade performance with explainability

D. Future Directions

CRISPR-Cas9 Optimization: Integrate DL for optimal gRNA sequence suggestion with off-target prediction. Federated Learning: Enable distributed hospital data training without privacy compromise, improving ethnic diversity coverage. Multimodal Fusion: Develop sophisticated fusion layer combining text/sequence outputs with learned attention mechanisms. Extended Coverage: Expand disorder training data and incorporate pharmacogenomic predictions for personalized treatment.

The NextGen AI Genomic Biomarker System represents significant progress toward solving precision medicine interpretation gaps, demonstrating AI can bridge phenotypic genotypic analysis while maintaining clinical transparency and accountability.

REFERENCES

[1]. B. Alipanahi et al., "Predicting sequence specificities by deep learning," *Nat. Biotechnol.*, vol. 33, no. 8, pp. 831–838, 2015.
[2]. L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[3]. H. J. Cordell, "Detecting gene-gene interactions," *Nat. Rev. Genet.*, vol. 10, no. 6, pp. 392–404, 2009.
[4]. J. Devlin et al., "BERT: Pre-training of Transformers," in *Proc. NAACL-HLT*, 2019.
[5]. D. Gunning, "Explainable AI (XAI)," DARPA, 2017.
[6]. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[7]. S. M. Lundberg and S. I. Lee, "Unified Approach to Model Predictions," in *NIPS*, vol. 30, 2017.
[8]. T.A. Manolio et al., "Missing heritability of complex diseases," *Nature*, vol. 461, pp. 747–753, 2009.
[9]. C. D. Manning and H. Schutze, ¨ *Foundations of Statistical NLP*. MIT Press, 1999.
[10]. R. Min et al., "LSTMs for Gene Regulation," in *IEEE BIBM*, 2017.
[11]. R. Poplin et al., "Universal variant caller using DNNs," *Nat. Biotechnol.*, vol. 36, no. 10, pp. 983–987, 2018.
[12]. C. Rudin, "Stop explaining black box ML models," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.
[13]. G. Salton and M. J. McGill, *Intro. to modern IR*. McGraw-Hill, 1983.
[14]. P. M. Visscher et al., "10 Years of GWAS Discovery," *Am. J. Hum. Genet.*, vol. 101, no. 1, pp. 5–22, 2017.
[15]. X. Wang et al., "Integrating clinical notes and genomic data," *BMC Med. Inform. Decis. Mak.*, vol. 21, pp. 1–13, 2021.
[16]. Wellcome Trust Consortium, "GWAS of 14,000 cases," *Nature*, vol. 447, pp. 661–678, 2007.

