International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Early Prediction of Spinal Muscular Atrophy (SMA) Using Machine Learning and Genomic Variant Analysis

Abhiraam S [1], Abhishek R [2], Darshan Sunil Kuranagi[3], D Sai Gagan[4], Manasa T P[5]

[1]*Abhiraam S, Dept. of Computer Science Engineering, Bangalore Institute of Technology, Karnataka, India*
[2]*Abhishek R, Dept. of Computer Science Engineering, Bangalore Institute of Technology, Karnataka, India*
[3]*Darshan Sunil Kuranagi, Dept. of Computer Science Engineering, Bangalore Institute of Technology, Karnataka, India*
[4]*D Sai Gagan, Dept. of Computer Science Engineering, Bangalore Institute of Technology, Karnataka, India*
[5]*Manasa T P, Dept. of Computer Science Engineering, Bangalore Institute of Technology, Karnataka, India*

Abstract - Spinal Muscular Atrophy (SMA) is a severe neuromuscular disease primarily caused by mutations in the Survival Motor Neuron 1 (SMN1) gene, leading to progressive motor neuron degeneration. Early diagnosis is critical as advanced therapies like Nusinersen are most effective before irreversible damage occurs. Current diagnostic practices, relying on symptom onset and manual genetic testing, often face delays. This paper proposes a machine learning based system for the early prediction of SMA by analyzing genomic variants. The system integrates clinical annotations from ClinVar with genetic sequence features extracted from FASTA files. We employ a hybrid feature extraction strategy, capturing sequence composition, splicing regulatory motifs, and SMN1/SMN2 paralog signatures. Ensemble learning algorithms, specifically Random Forest and XGBoost, are trained to classify variants as pathogenic or benign. The results demonstrate that the system effectively automates variant classification, offering a scalable and interpretable solution to support proactive healthcare decisions and early therapeutic intervention.

Keywords - Spinal Muscular Atrophy, Machine Learning, Genomic Variant Analysis, XGBoost, Random Forest, SMN1/ SMN2, Bioinformatics, Early Diagnosis.

I. INTRODUCTION

Spinal Muscular Atrophy (SMA) is a severe hereditary neuromuscular disease characterized by progressive degeneration of motor neurons in the spinal cord, leading to muscle weakness, atrophy, and respiratory complications. The disease affects approximately 1 in 10,000 live births globally. The molecular basis of SMA stems from homozygous mutations or deletions in the survival motor neuron 1 (SMN1) gene. These mutations result in the insufficient production of the essential SMN protein, which is vital for motor neuron survival. The advent of advanced treatments like Nusinersen, Risdiplam, and gene replacement therapies has demonstrated significant improvements in clinical outcomes. However, the effectiveness of these treatments is heavily dependent on administering them before irreversible motor neuron loss occurs. Current diagnostic practices, often reliant on symptom onset and standalone genetic tests, often result in diagnostic delays. To overcome these critical limitations, this project proposes a machine learning based early prediction system for SMA, utilizing comprehensive genetic information. The system aims to identify high-risk individuals even before the appearance of clinical symptoms by integrating annotated variant data from ClinVar with FASTA-formatted genetic sequences. Advanced classification algorithms, specifically Random Forest and XGBoost, are employed to extract predictive features and distinguish pathogenic variants associated with SMA from benign ones. This approach addresses the gap for an automated, scalable, and accurate prediction technique that combines both clinical and genetic biomarkers. The ultimate goal is to support proactive healthcare decisions, enabling earlier access to life-saving treatments.

SMA Type	Age of Onset	Typical SMN2 Copy Number	Motor Milestones	Prognosis
0	Prenatal/ at birth	1–2	Profound hypotonia, no milestone acquisition	Very poor, early neonatal death
I	<6 months	1–2	Never sit independently	High mortality in infancy without treatment
II	6–18 months	3	Sit but never walk independently	Survival into adolescence/adult hood with disability
III	>18 months	3–4	Stand and walk independently, later loss of ambulation	Variable, survival into adulthood
IV	Adult hood	4–5	Mild proximal weakness, normal early milestones	Near-normal life expectancy, mild symptoms

Table 1.1: *Feature categories used for SMA variant analysis.*

II. LITERATURE REVIEW

Traditional SMA diagnostic frameworks have long depended on the presence of observable clinical symptoms such as hypotonia, muscle weakness, and respiratory complications followed by confirmatory genetic testing focused on SMN1 and SMN2 copy number analysis. This symptom-first approach often leads to critical diagnostic delays, which significantly hinder therapeutic outcomes because motor neuron degeneration is irreversible and modern treatments like Nusinersen and Zolgensma provide maximum benefit when administered during pre symptomatic stages. Although newborn screening initiatives have improved early detection, the interpretation of rare or novel SMA-associated variants remains a major challenge for laboratories operating outside standardized workflows, limiting timely clinical decision-making [1]. To address these shortcomings, quantitative methods such as Quantitative Muscle Ultrasound (QMU) have been explored. QMU evaluates muscle deterioration through luminosity ratio measurements, showing strong correlation with clinical strength scores; however, it is constrained by small sample sizes, the exclusion of severe Type 1 cases, and its limited utility in preclinical prediction. Similarly, SMN protein quantification assays offer a biochemical indicator of disease severity, yet circulating SMN levels do not consistently reflect motor neuron-specific pathology, reducing their predictive reliability in early diagnosis [2]. Machine learning approaches have emerged as a promising supplement to clinical and genetic assessment, with recent studies applying supervised models to SMA prognosis and phenotype prediction. One such study utilized a Random Forest classifier to predict scoliosis progression using established clinical metrics such as HFMSE and CHOP INTEND, demonstrating the potential of ML to support ongoing patient management.

However, the study was constrained by a cohort of only 86 subjects and was applicable exclusively to already diagnosed patients. In another investigation, Weighted Gene Co-expression Network Analysis (WGCNA) combined with machine learning was applied to microarray datasets to identify biologically relevant gene modules and potential biomarkers. Despite the methodological innovation, the dataset included only sixteen samples, and the biologically important hub genes identified through the model lacked experimental validation, limiting clinical relevance. Overall, existing literature highlights meaningful progress in SMA monitoring and biomarker discovery while underscoring the need for scalable, predictive, and interpretable early-detection models capable of classifying variant pathogenicity before symptoms emerge [3].

A. Limitations of Existing Work

A review of the current literature reveals several significant limitations in existing approaches to SMA diagnostics and predictive modeling, underscoring the need for more effective and scalable solutions. Most machine learning efforts in SMA research emphasize prognosis rather than true early prediction, focusing on forecasting disease progression, severity, or treatment response in patients who have already been clinically diagnosed. Consequently, there is a notable lack of high-throughput computational models specifically designed to classify the pathogenicity of SMA associated genetic variants before symptom onset. Additionally, the datasets used in many published studies are small and highly constrained, microarray samples. Such limited cohort sizes restrict the robustness and generalizability of the resulting models, making them inadequate for population scale screening or for interpreting the broad spectrum of rare variants encountered in real-world genetic datasets. Another critical limitation lies in the absence of SMA specific feature engineering within current variant pathogenicity predictors. Most existing tools are generalized models trained across the entire human genome and therefore fail to incorporate key domain-specific genetic characteristics relevant to SMA. These models often overlook essential features such as SMN1 and SMN2 copy number variations, the molecular consequences of the c.859G>C single nucleotide variant, or the disruption of canonical and cryptic splice sites—factors that have direct implications for SMN protein expression. Furthermore, interpretability remains insufficient in many ML-based approaches. While some studies claim to integrate interpretable frameworks, the resulting predictions often rely on complex feature sets that fail to provide a biologically intuitive explanation. For clinical acceptance, prediction outputs must clearly articulate the underlying mechanisms—such as whether a variant exerts its effect through splicing disruption, altered protein stability, or another pathogenic process—to support clinician trust and decision-making.

Finally, current computational tools lack the scalability and accessibility required for real-world clinical deployment. Many of the existing systems function as academic prototypes without the automated data-processing capabilities necessary to handle raw VCF files produced by high-throughput sequencing platforms. Their limited computational efficiency and absence of user-friendly interfaces restrict usage to specialized bioinformatics experts, preventing integration into routine clinical diagnostics and g enetic counseling workflows. Collectively, these limitations highlight the need for an improved diagnostic framework that is predictive, interpretable, computationally scalable, and accessible to both clinicians and researchers.

III. METHODOLOGY

The proposed system for the early prediction of Spinal Muscular Atrophy (SMA) adopts a structured, multi stage machine learning framework that integrates extensive genomic data acquisition with biologically informed feature extraction and ensemble based classification techniques. The overall methodology is organized into four key phases: data collection and preprocessing, feature engineering, model development, and performance evaluation.

A. Data Acquisition and Preprocessing

The system collects genomic data from two main sources: ClinVar and the National Center for Biotechnology Information (NCBI). ClinVar supplies curated variant information for spinal muscular atrophy (SMA)–related genes, primarily SMN1 and SMN2. Each variant record includes chromosomal location, nucleotide alteration, and clinical significance. Reference genomic sequences are obtained in FASTA format from the Ensembl database, using the GRCh38 human genome assembly as the reference coordinate system.

During preprocessing, the system applies standardized normalization procedures to resolve inconsistencies across datasets. Chromosome identifiers are harmonized to ensure compatibility between annotation sources. Variant Call Format (VCF) files are normalized through left alignment to maintain consistent representation of insertions and deletions. Multi-allelic variants are decomposed into individual biallelic records to facilitate downstream analysis. All nucleotide sequences are converted to uppercase, and any non-standard nucleotide symbols are replaced with ambiguous base indicators to preserve biological validity.

Each variant is validated against the reference genome to confirm that the reference allele matches the corresponding genomic position. Variant records that fail validation are flagged and removed from further analysis. For all validated variants, genomic sequence windows of multiple lengths are extracted around the variant site to capture local sequence features relevant to splicing regulation and structural stability.

B. Feature Extraction Framework

The feature extraction framework converts genomic sequences and variant annotations into numerical representations suitable for machine learning models. Features are derived from both nucleotide sequences and variant-level annotations to capture biochemical, splicing-related, and functional characteristics relevant to spinal muscular atrophy (SMA).

Sequence composition features describe fundamental biochemical properties of genomic regions. Guanine–cytosine (GC) content is calculated to reflect DNA stability and methylation potential. Sequence complexity is quantified using Shannon entropy, where lower values indicate repetitive or low-complexity regions. CpG dinucleotide enrichment is measured as the ratio of observed to expected CpG occurrences, providing insight into regulatory and methylation-associated patterns.

Required formulas:

GC Content:

$$GC = \frac{N_G + N_C}{L}$$

Shannon entropy:

$$H = -\sum_{i \in \{A,C,G,T\}} p_i \log_2 p_i$$

Observed/Expected CpG ratio

$$CpG_{O/E} = \frac{N_{CpG} \cdot L}{N_G \cdot N_C}$$

Splicing-related features capture regulatory mechanisms underlying SMA pathology. Canonical splice donor and acceptor motifs are identified at exon–intron boundaries using pattern matching. The frequency of exonic splicing enhancers (ESEs) and silencers (ESSs) is quantified using experimentally validated motif libraries. Polypyrimidine tract integrity is evaluated to estimate splice acceptor site strength. These features are essential for distinguishing functional SMN1 transcripts from alternatively spliced SMN2 transcripts.

Paralog-specific features enable discrimination between SMN1 and SMN2, which differ by a small number of nucleotides but exhibit distinct splicing behaviors. The characteristic cytosine-to-thymine substitution at position 6 of exon 7 is explicitly detected, as it disrupts an exonic splicing enhancer in SMN2. Additional paralog-specific sequence motifs are identified and quantified to estimate the likelihood of full-length protein production.

Variant-level features describe the structural and functional effects of individual mutations. Variants are categorized by type, including single nucleotide variants, insertions, deletions, and frameshifts. The genomic distance between variants and critical regulatory regions is computed. Variants previously reported as pathogenic in clinical databases are flagged to incorporate prior biological knowledge.

Feature Category	Representative Features	Purpose/Relevance
Sequence Composition	GC content, sequence entropy	Captures nucleotide composition and sequence complexity influencing splicing and RNA stability
Splicing Regulatory Features	Canonical splice sites, splice site strength, ESE/ESS motifs	Identifies variants that may disrupt normal pre-mRNA splicing
Paralog Differentiation	SMN1/SMN2 signature motifs, exon 7 critical position	Distinguishes functional SMN1 from splice-deficient SMN2
Variant Annotation	Variant type, coding consequence, proximity to splice sites	Characterizes structural and functional impact of variants
Functional Impact Scores	CADD score, SpliceAI score, conservation score	Quantifies predicted pathogenicity and evolutionary importance
Language Model Embeddings	Transformer-based sequence embeddings	Encodes local and long-range genomic context for classification

Table 3.1: *Clinical classification of SMA by onset, SMN2 copies, and prognosis.*

Functional annotations enrich the feature set with predictive scores derived from established bioinformatics tools. The Variant Effect Predictor annotates coding consequences such as missense, nonsense, and synonymous changes. Combined Annotation Dependent Depletion (CADD) scores provide a quantitative estimate of variant deleteriousness. SpliceAI predicts splicing disruption by computing delta scores for splice donor and acceptor site gains and losses. These features complement sequence-derived attributes by incorporating evolutionary conservation and functional constraint.

To capture long-range sequence dependencies not represented by handcrafted features, contextual embeddings are generated using a pre-trained transformer-based genomic language model. Input sequences are tokenized into overlapping 6-mer units and processed by the transformer to generate contextualized embeddings. Mean pooling across sliding windows produces fixed-length vectors suitable for integration with conventional features.

C. Machine Learning Pipeline

The machine learning pipeline integrates all extracted features into a single feature matrix comprising sequence composition metrics, splicing-related indicators, paralog-specific markers, variant-level annotations, and genomic language model embeddings. This unified representation captures complementary biological signals associated with SMA pathogenicity.

Data Preprocessing

Preprocessing steps are applied to ensure model robustness and to prevent information leakage. Missing values in numerical features are imputed using median values calculated exclusively from the training dataset. Features that directly encode clinical significance labels are removed to avoid learning trivial correlations. Categorical variables are transformed using one-hot encoding, and numerical features are standardized to remove scale-dependent bias.

Required formula (z-score normalization):

$$Z = \frac{x-\mu}{\sigma}$$

where x is the feature value, $\mu$is the mean, and σ is the standard deviation computed from the training set.

Dataset Partitioning

The dataset is divided into training and testing subsets using stratified random sampling with an 80:20 split. Stratification preserves the relative proportions of pathogenic and benign variants across both subsets, which is critical for rare disease classification tasks.

Classification Models

a) Random Forest

Random Forest is employed as an ensemble classifier composed of multiple decision trees trained on bootstrap samples of the training data. Each tree is constructed using a random subset of features, and final predictions are obtained through majority voting across the ensemble. Tree depth is constrained to mitigate overfitting, and class weights are adjusted to compensate for class imbalance.

b) Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting is used as a complementary ensemble approach in which decision trees are trained sequentially. Each tree focuses on correcting the residual errors of the previous ensemble. The model optimizes a regularized objective function that balances predictive accuracy and model complexity, using gradient-based optimization with second-order derivatives. The number of boosting iterations is selected based on validation performance to prevent overfitting.

Model Training and Optimization

Both classifiers are trained on the standardized training dataset. Hyperparameters are optimized using cross-validation to maximize generalization performance. The final trained models are serialized and stored for integration into the downstream prediction system.

D. Model Evaluation

Model performance is evaluated on a held-out test set using multiple complementary classification metrics. Overall accuracy measures the proportion of correctly classified variants. Precision quantifies the reliability of pathogenic predictions by measuring the fraction of predicted pathogenic variants that are truly pathogenic. Recall evaluates the model's sensitivity by measuring the proportion of actual pathogenic variants that are correctly identified. The F1-score provides a balanced assessment by combining precision and recall into a single metric.

The discriminative capability of the models is further assessed using the area under the receiver operating characteristic curve (AUC–ROC). This threshold-independent metric evaluates the ability of the classifier to rank pathogenic variants higher than benign variants across all possible decision thresholds.

Confusion matrices are used to visualize classification outcomes in terms of true positives, false positives, true negatives, and false negatives, enabling detailed analysis of error patterns. Feature importance analysis is performed to identify the most influential features contributing to model predictions, supporting interpretability and biological relevance.

To ensure robust performance estimation, five-fold stratified cross-validation is conducted. Stratification preserves class proportions in each fold, and evaluation metrics are averaged across folds to obtain stable estimates of generalization performance.

Accuracy:

$$\textbf{Accuracy} = \frac{TP+TN+FP+FN}{TP+TN}$$

Precision:

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

Recall:

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

F1-score:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

AUC–ROC:

$$AUC = \int_1^0 TPR(FPR)\,d(FPR)$$

where TPR is the true positive rate and FPR is the false positive rate.

IV. CONCEPTUAL AND ANALYSIS MODELING

The conceptual and analysis modeling phase establishes the architectural foundation and operational workflow of the SMA prediction system through structured representations that bridge biological requirements with computational implementation. This section presents the systematic decomposition of the system into manageable components, defines interaction patterns between subsystems, and establishes the data transformation pathways that enable genomic variant classification.

A. Data Flow Architecture

The complete operational workflow implements a sequential data transformation pipeline wherein each processing stage produces structured outputs that serve as canonical inputs for subsequent stages. This architectural pattern ensures data consistency, enables independent validation of intermediate results, and facilitates modular replacement of processing components without affecting downstream operations.

The workflow initiates with genomic data acquisition from two principal sources: curated variant annotations extracted from the ClinVar database and reference nucleotide sequences retrieved from NCBI repositories in FASTA format. The ClinVar dataset provides essential attributes including genomic coordinates, gene identifiers, nucleotide alterations, and clinical significance classifications that establish ground truth labels for supervised learning. FASTA sequences supply reference genetic information for SMA-associated genes, with particular emphasis on the SMN1 and SMN2 paralogs that exhibit critical functional differences despite high sequence similarity.

Initial processing operations parse and extract relevant attributes from both data streams, including positional information encoded in genomic coordinate systems and sequence-level alterations that distinguish pathogenic variants from benign polymorphisms. Reference gene sequences undergo computational retrieval and processing through established bioinformatics utilities that ensure consistent formatting and coordinate system alignment. These parallel data streams converge during feature integration, where sequence-level characteristics are merged with variant-specific annotations to generate a comprehensive biological representation suitable for machine learning analysis.

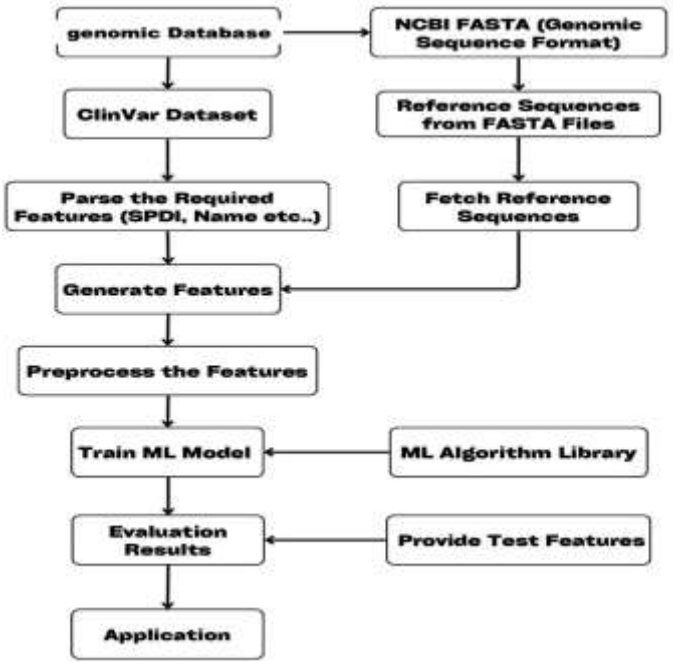

Figure 4.1: *Data Flow Diagram*

The feature engineering phase applies domain-specific transformations that capture both sequence composition properties and biologically validated markers associated with SMA pathogenesis. Sequence-based features quantify nucleotide composition patterns, including GC content ratios, dinucleotide frequencies, and entropy measures that reflect local sequence complexity. Variant-level features encode mutation classifications, predicted functional impacts, and splice site disruption potentials derived from established annotation frameworks. Critical SMA-specific signatures receive explicit representation, including the canonical SMN2 exon 7 nucleotide transition that distinguishes functional SMN1 alleles from the disease-modifying SMN2 paralog.

Following feature extraction, the structured dataset undergoes preprocessing operations that address missing values, normalize feature scales, and apply dimensionality reduction techniques when appropriate. This preprocessing phase ensures numerical stability during model training and prevents scale-dependent features from dominating learned decision boundaries. The resulting feature matrix serves as input to the machine learning module, where ensemble classification algorithms including Random Forest and XGBoost undergo training using standard optimization procedures.

Model evaluation employs rigorous validation protocols that assess predictive performance across multiple complementary metrics. Accuracy quantifies overall classification correctness, while precision and recall capture the system's ability to correctly identify pathogenic variants while minimizing false positive predictions. The F1-score provides a harmonic balance between precision and recall, ensuring that model optimization does not sacrifice one metric to artificially inflate the other. Area under the receiver operating characteristic curve serves as a threshold-independent measure of discriminative ability, reflecting the model's capacity to correctly rank variants by pathogenicity probability across all possible decision boundaries.

B. Architectural Design and Component Organization

The system architecture implements a layered design pattern that separates concerns across six distinct functional layers, each responsible for specific aspects of the genomic analysis workflow. This stratified organization promotes maintainability through clear interface definitions, enables independent testing of isolated components, and supports incremental enhancement without requiring wholesale architectural redesign.

The data acquisition layer orchestrates retrieval of genomic information from multiple heterogeneous sources, implementing robust error handling for network failures and data format inconsistencies. This layer normalizes chromosome nomenclature variations, resolves coordinate system discrepancies between reference assemblies, and validates data integrity through checksum verification. All acquired data undergoes schema validation before persistence to ensure downstream components receive consistently formatted inputs.

The preprocessing and normalization layer applies standardized transformations that convert raw genomic data into canonical representations suitable for biological analysis. FASTA sequence processing employs established bioinformatics libraries to parse sequence headers, extract nucleotide strings, and compute basic composition statistics. Variant call format files undergo normalization procedures that left-align insertion-deletion polymorphisms, decompose

multi-allelic sites into independent records, and validate reference allele concordance against the reference genome. These normalization operations eliminate technical artifacts introduced during sequencing and variant calling, ensuring that biological signal rather than technical noise drives subsequent classification decisions.

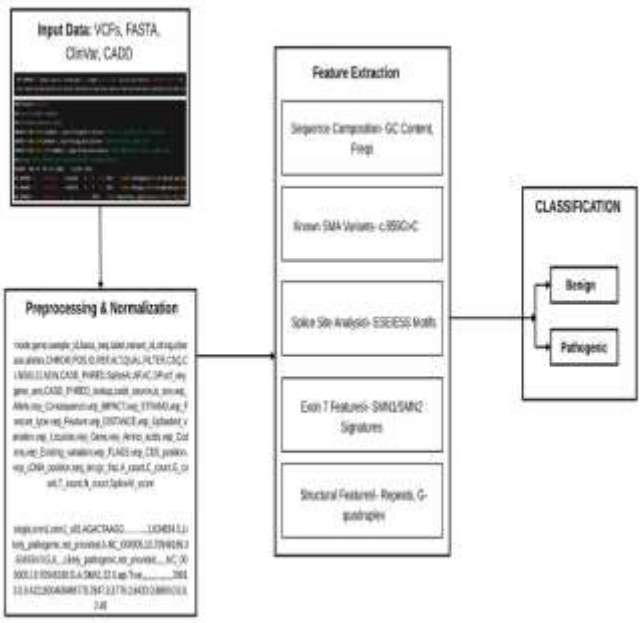

Figure 4.2: *Architecture Diagram of Early Prediction of SMA Using Machine Learning*

The feature extraction layer implements the critical translation from biological sequences to numerical representations amenable to machine learning algorithms. This layer computes sequence composition features including nucleotide frequencies, GC content ratios, and Shannon entropy measures that capture local sequence complexity. Splice site analysis identifies canonical donor and acceptor motifs, quantifies splice site strength using position weight matrices, and detects exonic splicing enhancer and silencer sequences that modulate exon inclusion efficiency. SMA-specific feature extraction explicitly encodes known pathogenic variants, computes SMN1 versus SMN2 discriminating signatures centered on the exon 7 critical nucleotide, and estimates copy number proxies through k-mer abundance analysis.

Genomic language model integration provides complementary sequence representations that capture long-range dependencies and subtle regulatory patterns not explicitly encoded in handcrafted features. Pre-trained transformer models receive k-mer tokenized sequence windows as input, generating high-dimensional embedding vectors that encode learned biological semantics. These contextual embeddings augment traditional feature sets with representations that reflect patterns implicit in large genomic corpora, potentially identifying predictive signals not captured by manually designed features.

The modeling layer implements ensemble classification through parallel deployment of complementary machine learning algorithms. Random Forest classifiers leverage bootstrap aggregation of decision trees to reduce prediction variance and provide robust performance across diverse feature distributions. XGBoost employs gradient boosting to sequentially refine predictions through iterative error correction, capturing complex non-linear interactions between genomic features. Model training incorporates class weighting strategies to address inherent imbalances between pathogenic and benign variant frequencies in clinical databases. Hyperparameter optimization employs cross-validation protocols that prevent overfitting while maximizing generalization performance on unseen variants.

The classification output layer generates structured prediction reports that communicate variant pathogenicity assessments along with supporting evidence and confidence metrics. Each prediction includes the binary classification decision, posterior probability estimates reflecting model certainty, and ranked feature contributions identifying which genomic properties most strongly influenced the classification outcome. Visualization components generate intuitive graphical representations of key findings, including feature importance plots, splice site disruption diagrams, and sequence context visualizations that facilitate clinical interpretation.

C. Detailed Component Interactions

The interaction between architectural components follows well-defined communication patterns that ensure data integrity and processing correctness throughout the analysis pipeline. Component interfaces specify explicit input and output contracts, including data types, required attributes, and acceptable value ranges. This formal specification enables independent component development and facilitates automated validation of interface compliance.

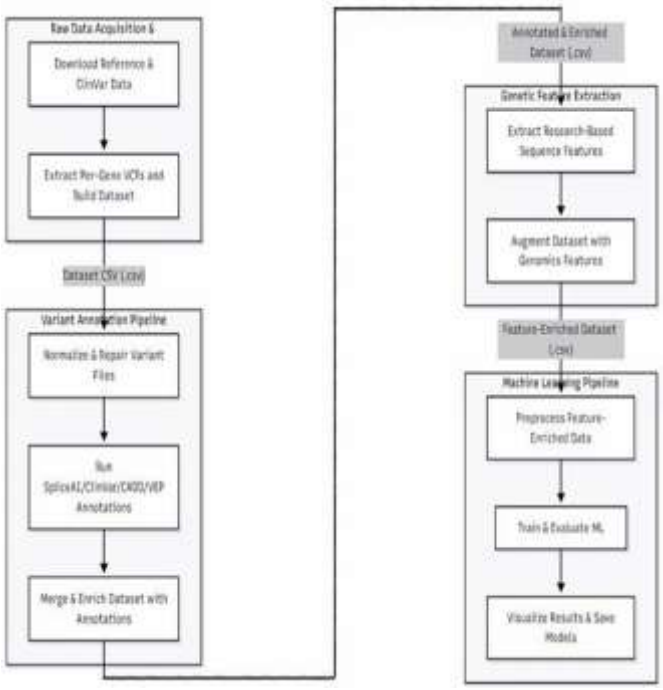

Figure 4.3: *Detailed Design*

The data retrieval module exposes interfaces for reference sequence fetching, variant extraction by genomic region, and coordinate system translation. Internal implementations handle authentication with external databases, implement retry logic for transient network failures, and cache frequently accessed reference sequences to reduce redundant network operations. The module returns structured objects containing sequence identifiers, nucleotide strings, genomic coordinates, and associated metadata required for downstream annotation operations.

The annotation pipeline module accepts normalized variant representations and orchestrates execution of multiple third-party bioinformatics tools including SpliceAI for splice impact prediction, Variant Effect Predictor for functional consequence annotation, and Combined Annotation Dependent Depletion for deleteriousness scoring. The module manages tool invocation through standardized command-line interfaces, monitors execution status, and aggregates results into unified annotation tables. Error handling implements graceful degradation strategies that allow pipeline continuation even when specific annotation sources become temporarily unavailable, logging missing annotations for subsequent manual review.

Feature extraction components consume annotated variant records and generate numerical feature vectors through application of transformation rules that encode biological properties. Sequence composition analyzers apply sliding window operations to compute local statistics, motif scanners employ regular expression matching to identify known regulatory elements, and structural analyzers detect repetitive regions and secondary structure forming sequences. All feature computations include explicit handling of edge cases including ambiguous nucleotides, incomplete annotations, and variants located near sequence boundaries.

The machine learning pipeline module implements the complete model lifecycle from data preprocessing through model serialization. Preprocessing operations apply imputation strategies for missing feature values, scale numerical features to zero mean and unit variance, and encode categorical variables through one-hot or ordinal encoding schemes. Model training employs stratified splitting to ensure representative class distributions in training and validation sets, implements early stopping to prevent overfitting, and persists trained model artifacts including learned parameters, feature scaling transformations, and label encoding mappings. Inference operations load serialized models, apply identical preprocessing transformations to new inputs, and generate predictions with associated confidence metrics derived from ensemble consensus or posterior probability estimates.

V. DISCUSSION

A. Interpretation of Experimental Results

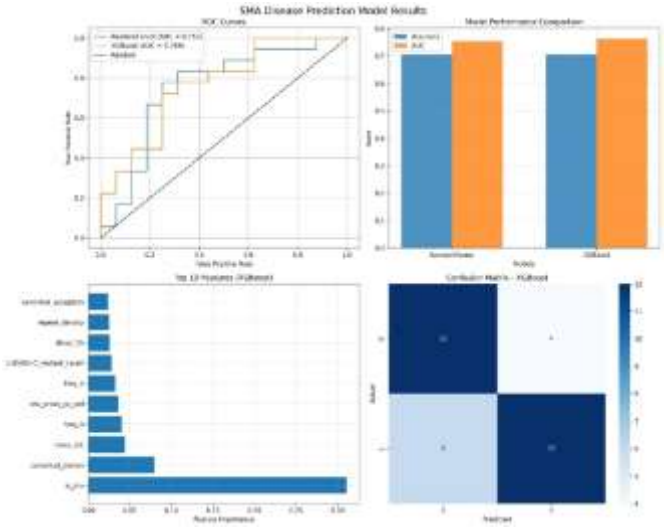

Figure 5.1: *SMA disease prediction model results*

The experimental evaluation of the proposed SMA prediction framework demonstrates the practical potential of machine learning techniques for rare genetic disease classification. The achieved accuracy of 71% using traditional ensemble models and 86% with the integration of a genomic language model indicates that computational approaches can effectively capture genetic patterns associated with Spinal Muscular Atrophy. While these results represent a significant improvement over random classification, they must be interpreted carefully in a clinical genomics context, where the consequences of incorrect predictions differ substantially between false positives and false negatives.

The improved performance observed with the DNABERT-enhanced pipeline supports the hypothesis that pre-trained genomic language models can identify latent sequence patterns not explicitly represented through handcrafted features. The observed 15% improvement in accuracy suggests that transformer-based models are capable of learning complex sequence dependencies, including regulatory elements and contextual signals, that are difficult to encode using traditional feature engineering methods. This observation is consistent with recent advances in natural language processing, where attention-based architectures have proven effective in modeling long-range dependencies in

sequential data.

Despite these improvements, the overall accuracy remains moderate, highlighting the intrinsic complexity of SMA variant classification. The SMN1 and SMN2 genes share nearly identical coding sequences, with disease relevance arising from subtle nucleotide differences that affect splicing efficiency rather than protein structure. This biological similarity imposes natural limits on predictive performance and suggests that further gains may require integration of additional data sources such as functional assays, clinical phenotypes, and family history information.

A comparison of ensemble classifiers shows that XGBoost marginally outperformed Random Forest across evaluation metrics. This difference can be attributed to XGBoost's gradient boosting strategy, which allows it to model complex non-linear relationships more effectively. The results indicate that the available dataset is sufficiently informative to benefit from this increased model capacity without severe overfitting.

B. Limitations and Sources of Error

Despite promising results, several limitations restrict the immediate clinical applicability of the proposed system. The most significant constraint is the limited size of the training dataset, which reflects the rarity of SMA and the scarcity of clinically validated variants. Small sample sizes increase prediction variability, reduce generalization performance, and limit the model's ability to learn rare mutation patterns.

Class imbalance between benign and pathogenic variants introduces additional bias toward the majority class. Although class weighting techniques were applied, some residual bias remains, potentially reducing sensitivity for detecting novel pathogenic variants.

Another limitation arises from reliance on computational splice prediction tools such as SpliceAI. While these tools are highly accurate, their prediction errors propagate through the classification pipeline, contributing to overall uncertainty.

The absence of direct copy number variation (CNV) detection is also a notable limitation, as homozygous deletions of SMN1 account for the majority of SMA cases. Sequence-based proxy features cannot fully substitute for dedicated CNV analysis methods, limiting the system's ability to detect this critical mutation class.

C. Clinical Implications

The results demonstrate the potential of automated variant classification systems as decision-support tools in clinical genetics. With an accuracy of 86%, the proposed framework can assist geneticists in prioritizing variants for further analysis, particularly those classified as variants of uncertain significance.

In newborn screening programs, automated pipelines offer the advantages of speed, consistency, and scalability. Early identification of SMA is especially critical given the availability of effective treatments that are most beneficial when administered before symptom onset.

The system's use of interpretable features and feature importance analysis addresses a key requirement for clinical adoption, namely transparency. By providing biologically meaningful explanations for predictions, the framework supports informed clinical decision-making.

However, the current performance level is insufficient for standalone diagnostic use. The model is better suited for triage

and prioritization, complementing rather than replacing expert clinical judgment. Furthermore, validation on independent clinical cohorts is required before real-world deployment.

VI. FUTURE WORK

The proposed system provides a strong basis for machine learning–based prediction of Spinal Muscular Atrophy (SMA); however, several enhancements can further improve its clinical relevance and performance. Future work should expand genomic coverage beyond SMN1 and SMN2 to include modifier genes and regulatory regions that influence disease severity. Integrating multi-omics data such as transcriptomics, proteomics, and epigenomics would offer a more comprehensive understanding of disease mechanisms. Advanced deep learning architectures and SMA-specific model fine-tuning could improve predictive accuracy. Incorporating longitudinal clinical data would enable modeling of disease progression over time. Predicting individual responses to available therapies could support personalized treatment planning. Privacy-preserving collaborative learning approaches may help overcome data scarcity while protecting patient confidentiality. External validation across diverse populations is essential for clinical reliability. Improved interpretability methods would enhance clinician trust in model outputs. Overall, these directions aim to evolve the system into a comprehensive precision medicine tool for SMA care.

VII. CONCLUSION

This work demonstrates that machine learning can be effectively applied for the early prediction of Spinal Muscular Atrophy (SMA) through genomic variant analysis. An end-to-end computational framework was developed that integrates curated genomic data, rigorous preprocessing, biologically informed feature extraction, and ensemble classification techniques to distinguish pathogenic variants from benign ones with clinically meaningful accuracy. The system incorporates established annotation tools such as VEP,

SpliceAI, and CADD to enrich sequence data with functional context, while explicitly modeling SMA-specific biological signals including SMN1–SMN2 paralog differentiation and exon 7 splicing characteristics. Ensemble classifiers using Random Forest and XGBoost achieved an accuracy of 71% on baseline features, which increased to 86% with the integration of DNABERT-based genomic language model embeddings, highlighting the benefit of contextualized sequence representations. The framework supports automated analysis of variants of uncertain significance and provides interpretable predictions through feature importance analysis, improving clinical transparency and usability. Although the system is currently limited by dataset size, class imbalance, and incomplete support for structural variants, it establishes a strong foundation for future extensions. With further

validation, integration of additional data modalities, and scalability enhancements, the proposed approach has the

potential to support early diagnosis and clinical decision-making for SMA.Overall, this study demonstrates that combining biological domain knowledge with modern machine learning techniques offers a practical and promising direction for advancing rare disease genomics and precision medicine.

ACKNOWLEDGMENT

We thank Bangalore Institute of Technology for research support, the open-source bioinformatics community (bcftools, samtools, SpliceAI, VEP), and ClinVar/NCBI for genomic resources.

REFERENCES

[1] D. Ramirez-Schrempp et al., "Interpretable ML model for anticipating SMA-associated scoliosis," Sci. Rep., vol. 14, no. 1, pp. 1-11, 2024.

[2] M. C. Pera et al., "Predictive models in SMA II natural history trajectories using ML," Neuromuscul. Disord., vol. 32, no. 5, pp. 345-353, May 2022.

[3] G. Coratti et al., "Machine-learning-based algorithm to predict therapeutic response in SMA," ResearchGate, 2023.

[4] A. El-Sayed et al., "Using network analysis and ML to identify genes implicated in SMA," J. Clin. Biomed. Res., vol. 4, no. 1, pp. 274537, 2023.

[5] P. Zaworski et al., "SMN protein measurement in whole blood for SMA clinical trials," PLOS ONE, vol. 11, no. 3, pp. e0150640, 2016.

[6] C. D. Wurster et al., "Neurofilaments in CSF show changes under nusinersen in SMA patients," Ther. Adv. Neurol. Disord., vol. 12, pp. 1-12, 2019.

[7] I. T. Zaharieva et al., "Plasma microRNAs respond to nusinersen in SMA patients," Ann. Clin. Transl. Neurol., vol. 9, no. 6, pp. 789-798, 2022.

[8] A. Yilmaz et al., "Neurofilament light chain as neuronal injury marker in SMA," Expert Rev. Mol. Diagn., vol. 17, no. 10, pp. 895-904, 2017.

[9] S. B. Rutkove et al., "Machine learning with EIM and QMU for SMA classification," Neurology, vol. 78, no. 14, pp. 1052-1058, 2012.

[10] J. S. Wu et al., "Quantitative ultrasound assessed SMA for classification," Ultrasound Med. Biol., vol. 36, no. 8, pp. 1234-1240, 2010.