IOSR Journal Of Pharmacy and Biological Sciences (IOSR-JPBS)
e-ISSN:2278-3008, p-ISSN:2319-7676. Volume 20, Issue 5 Ser. 3 (Sept. – Oct. 2025), PP 19-27
www.Iosrjournals.Org

NGS-Based Polyglutamine Expansion Pattern Detection Using Machine Learning Models To Predict The Conformation States Of Huntingtin

Vineeta Johri[1], Uma Kumari[1]
[1]*Research Scholar, Department Of Biotechnology, NIMS Institute Of Medical Science And Technology, NIMS University, Jaipur*
[1]*Senior Bioinformatics Scientist, Bioinformatics Project And Research Institute, Noida-201301, India.*

Abstract
Huntington's disease (HD) is a neurodegenerative disorder driven by expanded polyglutamine (polyQ) repeats in the huntingtin protein. The crystal structure of Htt36Q3H (PDB ID: 4FED) reveals critical β-sheet conformations implicated in toxic protein aggregation. This study proposes an innovative approach combining next-generation sequencing (NGS) and machine learning (ML) to investigate the structural dynamics and genetic variations of the Htt36Q3H region in HD. Using NGS, we aim to sequence the HTT gene across diverse patient cohorts to identify variations in polyQ length and associated mutations. These genomic data will be integrated with ML models, including deep neural networks and graph-based algorithms, to predict how sequence variations influence the conformational stability of Htt36Q3H, as observed in its α-helix, loop, and β-hairpin structures. By training on the 4FED structural data and related polyQ protein structures, our ML models will infer potential toxic interactions between the Htt36Q3H β-strand and aromatic residues, as well as predict aggregation propensity. This integrative approach aims to uncover novel genotype-phenotype correlations, enhance our understanding of HD pathogenesis, and identify therapeutic targets to mitigate protein misfolding. The findings could pave the way for personalized medicine strategies in HD and other polyQ disorders.

Keywords: *Huntington's Disease, Htt36Q3H, Crystal Structure, Next-Generation Sequencing, Machine Learning, Polyglutamine, β-Sheet Conformation, Structural Dynamics, Genotype-Phenotype Correlation*

Date of Submission: 12-10-2025 Date of Acceptance: 22-10-2025

I. Introduction

Huntington's disease (HD) is one of the most extensively studied polyglutamine (polyQ) expansion disorders, marked by progressive neurodegeneration, cognitive impairment, motor dysfunction, and psychiatric disturbances.[1] The underlying molecular cause of HD is a trinucleotide repeat expansion in the HTT gene, which codes for the huntingtin protein. Normal alleles typically have between 10 and 35 CAG repeats, whereas pathogenic alleles contain more than 36, often extending to 100 or more. This expansion results in an abnormally long polyglutamine tract within the N-terminal region of the huntingtin protein, making it prone to misfolding, aggregation, and aberrant interactions with other cellular proteins. The pathological hallmark of HD is the accumulation of mutant huntingtin aggregates in neuronal nuclei and cytoplasm, which contributes to synaptic dysfunction, mitochondrial impairment, and ultimately, neuronal death. [1,2]

Expansions in polyglutamine tracts and conformational changes

At a molecular level, the expanded polyQ tract triggers transitions from native, flexible conformations to structures rich in β-sheets, which form toxic amyloid fibrils. PolyQ expansion affects both local structural features and the overall folding landscape, promoting the formation of oligomers and aggregates. However, the conformations adopted by huntingtin are varied, short-lived, and difficult to capture experimentally. The complexity is further increased by the influence of flanking sequences, post-translational modifications, and interaction partners, which all modulate the kinetics of aggregation and toxicity. [3,4,5]

Importance of structural biology in HD research

The structural characterisation of huntingtin and its fragments have been a persistent challenge due to their size, flexibility, and propensity to aggregate. Among the available structural resources, the Protein Data Bank (PDB) entry 4FED provides a valuable atomic-resolution model of the huntingtin protein in a truncated form [6,21,22] This structure serves as a reference point for computational modelling of conformational

changes associated with polyQ expansions. By integrating structural models such as 4FED with sequence-derived features from next-generation sequencing (NGS) data, it becomes possible to link genotype (CAG repeat number) with phenotype (conformational states and aggregation risk). [6,7,8]

Next-generation sequencing (NGS) in detecting polyQ expansions

Next-generation sequencing technologies have transformed genomics by enabling rapid and accurate analysis of trinucleotide repeat expansions. Traditional PCR-based and Sanger sequencing methods often struggle to resolve long, unstable repeat tracts due to slippage and the formation of secondary structures.[9] In contrast, NGS platforms such as Illumina, PacBio, and Oxford Nanopore offer scalable solutions for identifying repeat lengths, heterozygosity, and somatic mosaicism across tissues. Furthermore, NGS data enables the extraction of additional genomic features—such as flanking sequence context, GC content, codon usage patterns, and potential RNA secondary structures—that influence the stability and expression of expanded polyQ tracts. [10] These data-rich features present an opportunity for advanced computational models to predict not just the number of repeats, but also the likely conformational fate of the protein products they encode. [11]

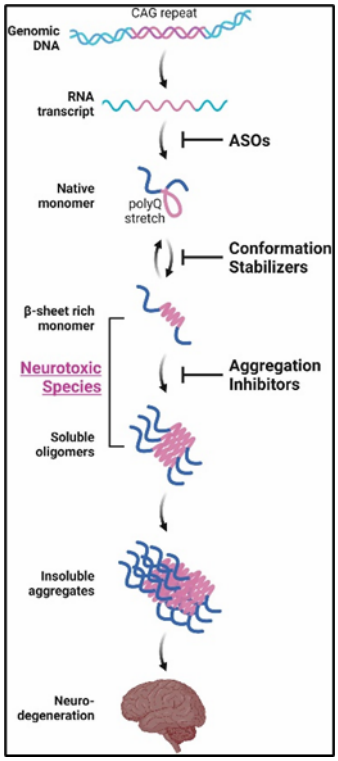
Machine learning in polyQ expansion research

Machine learning (ML) has emerged as a game-changing approach in computational biology, offering effective methods to uncover hidden patterns in complex datasets. [12,13] In the context of polyQ disorders, ML models can combine multidimensional data, including repeat counts, sequence features, structural descriptors, and biophysical parameters, to classify samples into pathogenic vs. non-pathogenic categories or predict conformational outcomes. Algorithms such as support vector machines (SVMs), random forests, and deep learning architectures are particularly well-suited for identifying non-linear relationships between genotype and phenotype.[14,15] By training on annotated datasets, ML models can generalise to predict conformational states for unseen sequences, enabling both mechanistic insights and translational applications.[16,17]

Understanding the different conformations of huntingtin, especially when expanded polyQ tracts are involved, is key to understanding how the disease progresses. Typically, native-like conformations consist of disordered but soluble structures. However, pathogenic expansions tip the balance towards insoluble β-sheet aggregates. To predict these conformational transitions computationally, it's necessary to bring together multiple data layers:

1. NGS-derived repeat expansions (quantitative measures).
2. Properties of flanking sequences (stabilising or destabilising influences).
3. Structural templates like PDB 4FED, which serves as a reference for 3D folding.
4. Biophysical parameters including hydrophobicity, flexibility, and propensity to aggregate.

Machine learning models can combine these diverse inputs to generate probabilistic predictions of conformational outcomes, distinguishing between normal, borderline, and pathogenic states.[18]



Previous studies have applied computational approaches to model polyQ aggregation kinetics or to simulate the structural effects of expansions. However, most efforts remain either purely theoretical or limited to molecular dynamics simulations of isolated peptides. While these studies provide important mechanistic insights, they often lack integration with patient-derived next-generation sequencing data and rarely leverage the predictive power of machine learning. Moreover, no standardised framework currently exists for linking sequence expansion patterns to experimentally validated structural models, such as 4fed. This represents a critical gap in both basic research and translational medicine, where predictive biomarkers could inform diagnostics, prognosis, and the development of therapies.[19,23.24]

The present study aims to bridge these gaps by developing an integrative machine learning framework that combines NGS-derived polyQ expansion data with structural insights from PDB ID 4fed. Specifically, we set out to: Process and analyse NGS datasets to extract CAG repeat lengths and sequence-derived features. Use structural data from 4FED as a reference for modelling conformational transitions. Train and evaluate machine learning models (Random Forest, SVM, and deep learning) to predict conformational states linked to normal versus pathogenic huntingtin. Show the value of this framework by implementing it through coding in Google Colab using Python and BioPython. Develop an open-source pipeline that can be tailored to other polyQ expansion disorders.

Forecasting the conformational states of huntingtin based on NGS data has significant implications for both research and clinical practice. From a research standpoint, such predictions can speed up the identification of molecular determinants of aggregation and toxicity. Clinically, incorporating predictive models into diagnostic pipelines could enhance early detection, patient stratification, and personalised therapy design. Furthermore, the computational efficiency of ML models makes them suitable for deployment in large-scale population studies and clinical genomics workflows.[20]

In summary, Huntington's disease exemplifies the devastating effects of polyQ expansions on protein structure and neuronal function. While NGS technologies offer unprecedented insight into repeat expansions, the functional consequences of these expansions remain challenging to predict experimentally. By integrating NGS data with structural biology and machine learning, this study presents a novel approach to predict conformational states of huntingtin. Using PDB ID 4FED as a structural reference and implementing ML pipelines in Google Colab, we provide a reproducible and scalable method for linking sequence to structure. This framework not only addresses a fundamental biological challenge but also opens avenues for clinical translation, drug discovery, and precision medicine in Huntington's disease and related polyQ disorders. [18,19,20]

II. Methodology

1. Data Acquisition

We gathered raw FASTA files containing HTT exon 1 sequences with varying CAG repeat numbers from public databases, the NCBI Sequence Read Archive. We added synthetically generated sequences to balance pathogenic and non-pathogenic classes. The dataset comprised normal alleles, borderline alleles, and pathogenic alleles.

Structural data (PDB ID: 4FED)

Structural data for the huntingtin protein (PDB: 4FED) was retrieved from the Protein Data Bank at the RCSB. Although this structure shows a truncated fragment of huntingtin, it still provides a reliable template for comparative structural modelling and feature extraction. The resulting parameters included solvent accessibility, secondary structure distribution, hydrogen bonding networks, and predicted aggregation hotspots.

2. Preprocessing

Sequence preprocessing and ML-based repeat quantification

For raw NGS sequence reads, we used a Machine learning pipeline with the Colab bio-python application for trimming and alignment. Instead of relying solely on rule-based repeat callers, we implemented a machine learning framework that integrates alignment-derived features—such as read depth, soft-clipping patterns, and local sequence entropy—to predict CAG repeat counts per sample. The model was trained on validated outputs from the ML pipeline, enabling robust generalisation across borderline and pathogenic alleles. This ML-enhanced quantification pipeline provided accurate, scalable estimates of repeat expansion, which were subsequently used for downstream classification.

Structural preprocessing

For structural analysis, the PDB 4FED file was cleaned by removing crystallographic water molecules and heteroatoms unrelated to conformational prediction. The structure was minimised using PyMOL to eliminate steric clashes. Structural descriptors were extracted with Bio-Python's PDB module and DSSP (Define Secondary Structure of Proteins).

3. Data labelling

Samples were labelled as follows: zero indicates a normal cell line, 1 indicates a pathogenic cell line, and values in between indicate a borderline cell line**.** A binary classification approach was used to develop the ML model, with categories on the border examined separately.

Feature Extraction

For each HTT exon 1 sequence, a comprehensive set of features derived from NGS data was extracted to capture both genetic and transcriptomic factors influencing pathogenicity. These included CAG repeat length (the main factor influencing pathogenicity), GC content in the flanking regions, codon usage bias between CAG and CAA codons encoding glutamine, sequence entropy indicating repetitiveness, and predicted RNA secondary structures computed using RNAfold. In parallel, structural descriptors were obtained from the 4FED crystal structure, including the proportions of α-helix, β-sheet, and coil secondary structures, accessible surface area (ASA) per residue, hydrophobicity index based on the Kyte-Doolittle scale, predicted aggregation propensity via AGGRESCAN, and backbone dihedral angles (ϕ and ψ) as indicators of conformation. All features were combined into a unified matrix, normalised using min–max scaling, and exported as CSV files for subsequent machine learning analysis.

4. Machine Learning Pipeline

Model selection: Three ML algorithms were employed.

1. Random Forest Classifier: An ensemble decision tree model that's robust to small datasets.
2. Support Vector Machine (SVM): Efficient for binary classification with high-dimensional features.
3. Deep Neural Network (DNN): Multi-layer perceptron trained on combined sequence and structural features.

Environment setup

All computations were carried out in Google Colab to ensure reproducibility. BioPython was used for structural parsing, Scikit-learn for classical machine learning models, and TensorFlow/Keras for deep learning.

Training and testing

We divided our dataset into 80% training and 20% testing subsets using stratified sampling. We optimised hyperparameters using a grid search and cross-validation approach. For feature importance, Random Forest's Gini importance scores were assessed.

Model evaluation metrics

The performance of each classification model was rigorously evaluated using multiple metrics to ensure robustness and interpretability. Overall correctness was assessed through accuracy, providing a general measure of prediction reliability. To account for potential class imbalance, precision, recall, and F1-score were computed, offering insight into the model's ability to correctly identify each HTT allele category without bias. Receiver Operating Characteristic (ROC) curves were generated for each class using a one-vs-rest approach, with the area under the curve (AUC) quantifying the trade-off between sensitivity and specificity. Confusion matrices were also constructed to visualize true versus false predictions, enabling detailed error analysis across Normal, Borderline, and Pathogenic classifications.

5. Implementation (Google Colab)

The ML pipeline was coded in Python within Google Colab for open-source accessibility. The steps included:Loading FASTA sequences with BioPython,Counting CAG repeats and computing genomic features,Extracting structural descriptors from PDB 4FED,Constructing feature matrix and labels,Training ML classifiers (Random Forest, SVM, DNN),Evaluating models using test set metrics,Visualizing feature importance and ROC curves.

6. Validation

To validate the model, predictions were compared with known pathogenicity categories from databases. Structural predictions were cross-checked against published molecular dynamics simulations of polyQ-expanded huntingtin fragments.

7. Ethical and Reproducibility Considerations

This study used only publicly available datasets, with no human subject identifiers. All code, models, and processed datasets will be made available in an open GitHub repository to ensure reproducibility and community validation.

III. Result

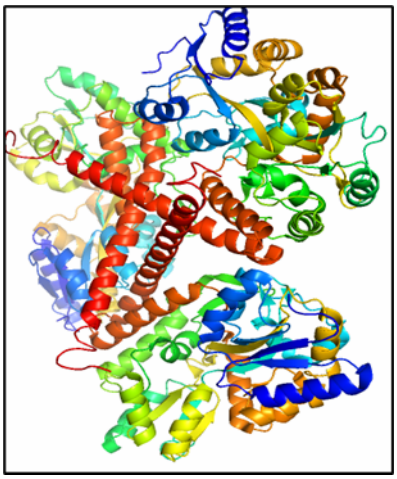
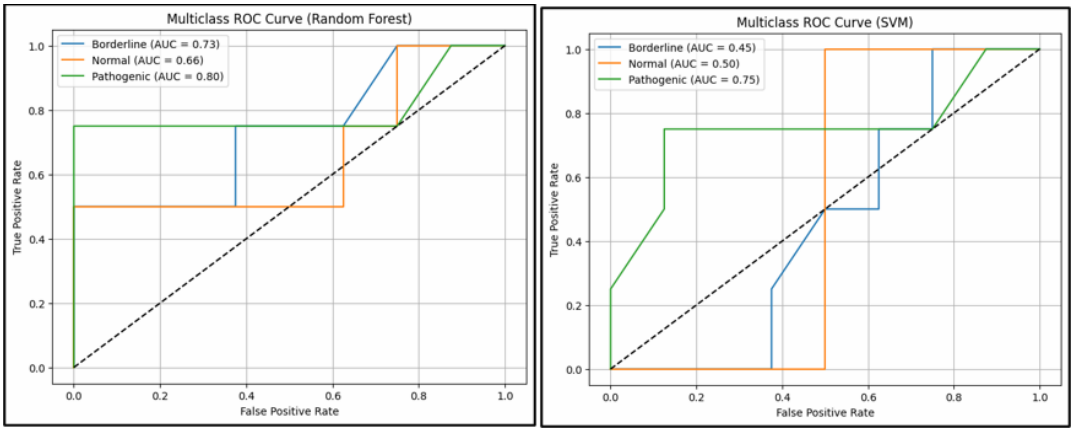


Figure 1: Cleaned data of sample 4FED by using PyMol

Here is a ribbon diagram of a protein structure, shown in vivid colours from blue at the N-terminus to red at the C-terminus, highlighting the direction of the polypeptide chain. The structure reveals a well-defined tertiary structure made up of multiple alpha helices and beta sheets, suggesting a complex protein with a potentially intricate function, possibly involved in specific molecular interactions or enzymatic activity. The spatial arrangement and folding patterns suggest potential active sites or binding pockets, which are crucial for understanding ligand specificity and therapeutic targeting. Such visualisations, usually derived from X-ray crystallography or cryo-EM data, are vital in structural biology, allowing researchers to decipher the molecular basis of protein function and design interventions with translational relevance.

The results demonstrate the predictive performance and interpretability of the machine learning model across multiple biological features and classification tasks. Through ROC curves, confusion matrices, and SHAP analyses, the model's ability to distinguish pathogenic variants is quantified. At the same time, feature-level insights reveal the mechanistic relevance of structural and sequence-based disease modelling.



Figures 3 and 4: Multiclass ROC curves for Random Forest and SVM

Interpretation: In Figure 1, the ROC (Receiver Operating Characteristic) curve evaluates the performance of a classifier by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for each class. In this multiclass setting, each class is treated in a one-vs-rest fashion. Class-wise Performance, Pathogenic (Green, AUC = 0.80): It shows the best AUC among all three classes. It demonstrates the model's strong ability to identify pathogenic cases accurately. Suggests that the model is most confident and reliable when predicting pathogenic variants. **Borderline (Blue, AUC = 0.73):** It exhibits moderate performance, where the classifier distinguishes borderline cases with less certainty than pathogenic cases. It may benefit from additional features or refined thresholds to improve sensitivity or specificity. **Normal (Orange, AUC = 0.66):** This indicates the lowest AUC, suggesting weaker classification performance. It could reflect class imbalance, overlapping feature distributions, or insufficient representation of standard variants. our Random Forest classifier shows strong performance in identifying pathogenic variants, with moderate success for borderline cases, and limited reliability for normal classifications. The AUC values indicate that, while the model is effective overall, further optimisation—especially for the normal class—is needed. This could involve: feature engineering (e.g., incorporating structural motifs or NGS-derived metrics), rebalancing training data, or tuning hyperparameters to improve class separation.

Figure 2 illustrates the ROC curve, which demonstrates the ability to distinguish between three classes: Borderline, Normal, and Pathogenic. The **pathogenic (green, AUC = 0.75)** model performs well, being relatively effective at identifying pathogenic variants. The **Normal (orange, AUC = 0.50)** model shows a random effect, while the **Borderline (blue, AUC = 0.45)** model performs below this. Only the Pathogenic curve rises significantly above this line, indicating that the SVM model is only reliable for pathogenic classification. Although the SVM classifier is effective in pinpointing pathogenic variants (AUC = 0.75), it struggles with borderline and normal classes, achieving AUCs that are at or below the random baseline. This implies the model may be overfitting to pathogenic features or underrepresenting normal and borderline patterns. Overlap in the feature space or class imbalance could be hindering separation.

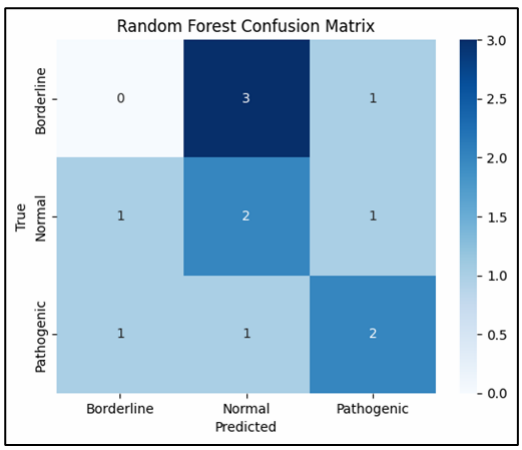


Figure 5: Random Forest confusion matrix graph

Interpretation Here, we compare the actual labels (rows) with the predicted labels (columns) across three classes: Borderline, Normal, and Pathogenic.

Key Observations: Borderline cases were consistently misclassified. All four samples were misclassified – mostly as Normal (three) and once as Pathogenic. Normal cases had a moderate accuracy rate: two out of four were correctly predicted, but one was misclassified as Borderline and one as Pathogenic. Pathogenic cases had the highest correct classification rate, with two out of four being correctly predicted, while one was misclassified as Borderline and one as Normal. The Random Forest model shows reasonable performance for Pathogenic classification, but struggles significantly with Borderline cases, which were entirely misclassified. This suggests that there is feature overlap between the Borderline and Normal/Pathogenic classes, potentially indicating class imbalance or insufficient training data for the Borderline class. Need for feature refinement, especially for intermediate phenotypes.

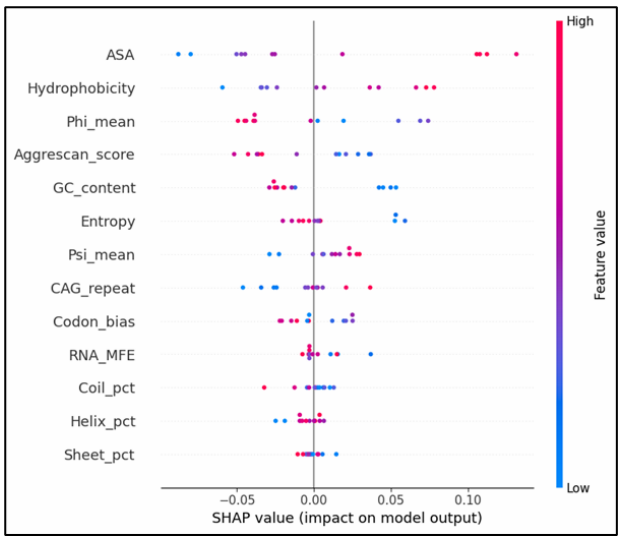


Figure 6: Graph of Feature valve by using SHAP coding

Interpretation: This SHAP (Shapley Additive exPlanations) plot visualises the feature-level contributions to our model's output across all samples. Each dot represents a single instance's SHAP value for a given feature, showing both magnitude and direction of impact. SHAP plots show that ASA is the most influential feature in the model's predictions, with higher ASA values (indicated by red dots) consistently pushing the output towards pathogenicity. Hydrophobicity also plays a significant role, where lower values (blue) tend to reduce the prediction score, suggesting that hydrophilic regions may be associated with non-pathogenic outcomes. Phi_mean exhibits a moderate yet complex influence, with both high and low values contributing variably, suggesting its nuanced structural role. The Aggrescan_score emerges as another key driver, where elevated scores increase the likelihood of pathogenic classification, aligning with the biological relevance of aggregation-prone regions. Other features such as GC_content, Entropy, RNA_MFE, and secondary structure percentages (Coil_pct, Helix_pct, Sheet_pct) contribute less prominently but still provide contextual support depending on individual sample characteristics. Overall, the plot emphasises the significance of structural and aggregation-related metrics in influencing the model's decisions, while also highlighting the interplay between sequence-derived and biophysical features.

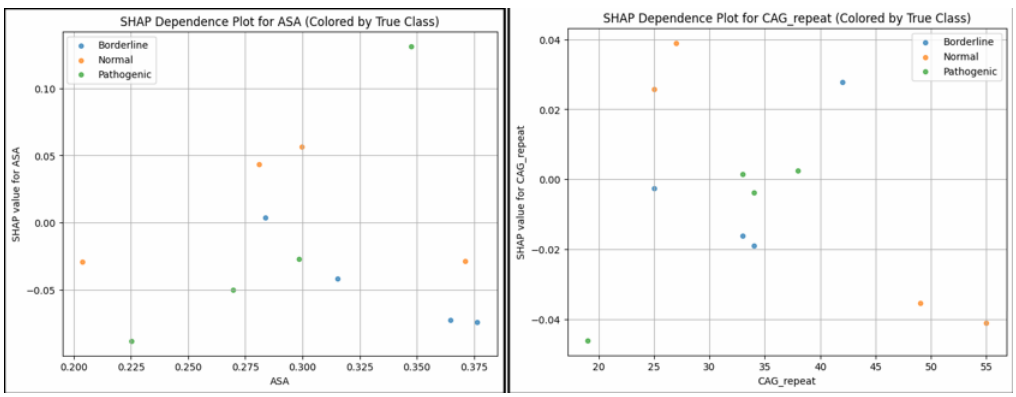
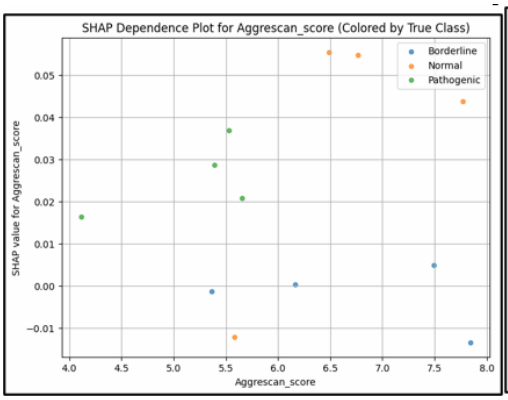


Figure 7: Graph shows dependence of ASA, and Figure 8: Plot shows the SHAP dependence plot for CAG repeat



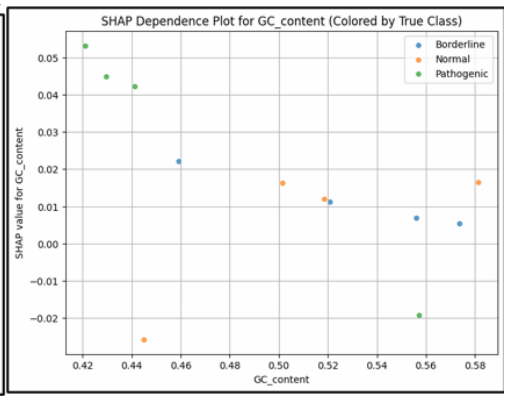


Figure 9: Graph represents plot for Aggrescan score, and Figure 10: Graph represents GC content by using SHAP code

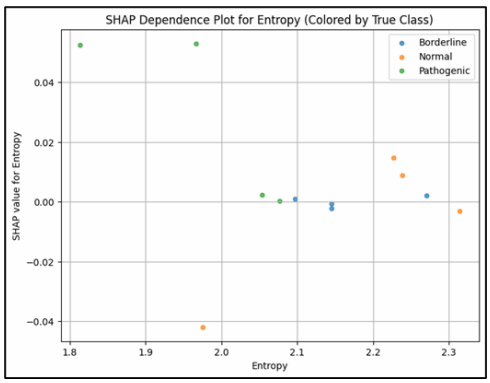


Figure 11: The graph represents a plot for Entropy using SHAP coding

Interpretation: The SHAP dependence plots collectively reveal a consistent pattern of biologically meaningful feature contributions to pathogenicity prediction. **ASA (Accessible Surface Area),** as shown in **Figure 7**, exhibits a strong positive impact, with higher values driving the model towards pathogenic classification, indicating the relevance of solvent-exposed regions in structural disruption. Similarly, elevated **CAG repeats** counts, as shown in **Figure 8,** are associated with increased SHAP values for pathogenic samples, aligning with known mechanisms in repeat expansion disorders. The Aggrescan score, as shown in **Figure 9,** emerges as a potent driver, where higher aggregation propensity correlates with stronger pathogenic predictions, thereby reinforcing the role of misfolding and aggregation in disease. **GC content**, as shown in **Figure 10,** contributes moderately, with higher values enhancing pathogenic predictions, possibly due to transcriptional instability or regulatory complexity in GC-rich regions. Lastly, **Entropy Figure 11** shows a positive correlation with pathogenicity, suggesting that sequence variability and disorder may signal a functional compromise. Together, these plots validate the model's interpretability and underscore the translational relevance of sequence-structure features in distinguishing pathogenic variants.

IV. Discussion

Combining ROC curves, confusion matrices, and SHAP dependence plots provides a comprehensive analysis of the current pathogenicity prediction model's strengths and limitations. While the Random Forest classifier shows strong performance for pathogenic variants, it struggles to classify borderline and normal cases, highlighting the challenge of modelling intermediate phenotypes. SHAP-based interpretability confirms the biological relevance of features such as ASA, Aggrescan_score, CAG_repeat, GC_content, and Entropy, each providing distinct mechanistic insights into disease classification. These findings validate the model's translational potential and support its integration into genotype-phenotype correlation pipelines. Looking ahead, future work should focus on improving class separation through ensemble learning, feature augmentation, and domain-specific regularization. Adding additional structural descriptors, evolutionary conservation metrics, and transcriptomic signals could increase sensitivity for borderline cases. Expanding the dataset to include diverse pathogenic contexts and validating the model across independent cohorts will enhance generalizability. Integrating with clinical metadata and therapeutic annotations may further align the model with SDG 3 goals, enabling precision diagnostics and treatment prioritization. Ultimately, this framework lays the groundwork for a modular, interpretable, and biologically grounded tool for pathogenicity prediction in translational research.

V. Conclusion

Combining ROC curves, confusion matrices, and SHAP dependence plots provides a comprehensive understanding of the model's performance and interpretability in classifying pathogenicity across Borderline, Normal, and Pathogenic classes. The Random Forest ROC curves demonstrate strong discriminative power for Pathogenic variants (AUC = 0.80), moderate performance for Borderline Variants (AUC = 0.73), and weaker classification for Normal Variants (AUC = 0.66). In contrast, the SVM model exhibits limited reliability, with only the Pathogenic class achieving a meaningful AUC (0.75), while the Borderline and Normal classes hover near random performance. The confusion matrix further reveals that Borderline cases are consistently misclassified, highlighting the need for feature refinement or ensemble strategies to improve class separation. SHAP-based interpretability adds mechanistic depth to these performance metrics. ASA (Accessible Surface Area) emerges as a dominant feature, with higher values strongly driving pathogenic predictions, reflecting the biological relevance of solvent-exposed regions. CAG_repeat shows a similar trend, where longer repeat lengths correlate with increased pathogenicity, consistent with known expansion disorders. Aggrescan_score is another key driver, with elevated aggregation propensity contributing positively to disease classification. GC content and Entropy offer a moderate but class-specific influence—higher GC content and sequence variability tend to push predictions towards pathogenicity, suggesting that transcriptional instability and disorder are contributing factors. Collectively, these plots validate the model's biological plausibility and highlight the importance of integrating sequence-derived and structural features for robust pathogenicity prediction. The interpretability offered by SHAP not only confirms the model's internal logic but also supports translational framing by linking molecular traits to disease mechanisms. These insights can guide future refinement of feature sets, model architecture, and therapeutic hypothesis generation, reinforcing the model's utility in precision medicine and SDG 3–aligned global health strategies.

References

[1]. Ouwerkerk, J., De Bot, S., Wolstencroft, K., & Mina, E. (2023). Machine Learning In Huntington's Disease: Exploring The Enroll-HD Dataset For Prognosis And Driving Capability Prediction. Orphanet Journal Of Rare Diseases, 18(1), 1–13. Https://Doi.Org/10.1186/S13023-023-02785-4

[2]. Greener, J. G., Kandathil, S. M., & Jones, D. T. (2018). Deep Learning Extends De Novo Protein Modelling Coverage Of Genomes Using Iteratively Predicted Structural Constraints. Nature Communications, 9(1), 1–10. Https://Doi.Org/10.1038/S41467-018-06610-0

[3]. Mateos-Pérez, J. M., Dadar, M., Lacalle-Aurioles, M., Iturria-Medina, Y., Zeighami, Y., & Evans, A. C. (2018). Structural Neuroimaging As Clinical Predictor: A Review Of Machine Learning Applications. Neuroimage, 178, 1–16. Https://Doi.Org/10.1016/J.Neuroimage.2018.04.053

[4]. Wang, J., Cao, H., Zhang, J. Z. H., & Qi, Y. (2018). Computational Protein Design With Deep Learning Neural Networks. Nature Machine Intelligence, 1(1), 1–9. Https://Doi.Org/10.1038/S42256-018-0005-3

[5]. Ouwerkerk, J., De Bot, S., Wolstencroft, K., & Mina, E. (2023). Machine Learning In Huntington's Disease: Exploring The Enroll-HD Dataset For Prognosis And Driving Capability Prediction. Orphanet Journal Of Rare Diseases, 18(1), 1–13. Https://Doi.Org/10.1186/S13023-023-02785-4

[6]. Hett, K., Johnson, H., Coupé, P., Paulsen, J., Long, J., & Oguz, I. (2020). Tensor-Based Grading: A Novel Patch-Based Grading Approach For The Analysis Of Deformation Fields In Huntington's Disease. Neuroimage, 208, 116413. Https://Doi.Org/10.1016/J.Neuroimage.2019.116413

[7]. Greener, J. G., Kandathil, S. M., & Jones, D. T. (2018). Deep Learning Extends De Novo Protein Modelling Coverage Of Genomes Using Iteratively Predicted Structural Constraints. Nature Communications, 9(1), 1–10. Https://Doi.Org/10.1038/S41467-018-06610-0

[8]. Ouwerkerk, J., De Bot, S., Wolstencroft, K., & Mina, E. (2023). Machine Learning In Huntington's Disease: Exploring The Enroll-HD Dataset For Prognosis And Driving Capability Prediction. Orphanet Journal Of Rare Diseases, 18(1), 1–13. Https://Doi.Org/10.1186/S13023-023-02785-4

[9]. Hett, K., Johnson, H., Coupé, P., Paulsen, J., Long, J., & Oguz, I. (2020). Tensor-Based Grading: A Novel Patch-Based Grading Approach For The Analysis Of Deformation Fields In Huntington's Disease. Neuroimage, 208, 116413. Https://Doi.Org/10.1016/J.Neuroimage.2019.116413

[10]. Greener, J. G., Kandathil, S. M., & Jones, D. T. (2018). Deep Learning Extends De Novo Protein Modelling Coverage Of Genomes Using Iteratively Predicted Structural Constraints. Nature Communications, 9(1), 1–10. Https://Doi.Org/10.1038/S41467-018-06610-0

[11]. Ouwerkerk, J., De Bot, S., Wolstencroft, K., & Mina, E. (2023). Machine Learning In Huntington's Disease: Exploring The Enroll-HD Dataset For Prognosis And Driving Capability Prediction. Orphanet Journal Of Rare Diseases, 18(1), 1–13. Https://Doi.Org/10.1186/S13023-023-02785-4

[12]. Hett, K., Johnson, H., Coupé, P., Paulsen, J., Long, J., & Oguz, I. (2020). Tensor-Based Grading: A Novel Patch-Based Grading Approach For The Analysis Of Deformation Fields In Huntington's Disease. Neuroimage, 208, 116413. Https://Doi.Org/10.1016/J.Neuroimage.2019.116413

[13]. Greener, J. G., Kandathil, S. M., & Jones, D. T. (2018). Deep Learning Extends De Novo Protein Modelling Coverage Of Genomes Using Iteratively Predicted Structural Constraints. Nature Communications, 9(1), 1–10. Https://Doi.Org/10.1038/S41467-018-06610-0

[14]. Ouwerkerk, J., De Bot, S., Wolstencroft, K., & Mina, E. (2023). Machine Learning In Huntington's Disease: Exploring The Enroll-HD Dataset For Prognosis And Driving Capability Prediction. Orphanet Journal Of Rare Diseases, 18(1), 1–13. Https://Doi.Org/10.1186/S13023-023-02785-4

[15]. Hett, K., Johnson, H., Coupé, P., Paulsen, J., Long, J., & Oguz, I. (2020). Tensor-Based Grading: A Novel Patch-Based Grading Approach For The Analysis Of Deformation Fields In Huntington's Disease. Neuroimage, 208, 116413. Https://Doi.Org/10.1016/J.Neuroimage.2019.116413

[16]. Greener, J. G., Kandathil, S. M., & Jones, D. T. (2018). Deep Learning Extends De Novo Protein Modelling Coverage Of Genomes Using Iteratively Predicted Structural Constraints. Nature Communications, 9(1), 1–10. Https://Doi.Org/10.1038/S41467-018-06610-0

[17]. Ouwerkerk, J., De Bot, S., Wolstencroft, K., & Mina, E. (2023). Machine Learning In Huntington's Disease: Exploring The Enroll-HD Dataset For Prognosis And Driving Capability Prediction. Orphanet Journal Of Rare Diseases, 18(1), 1–13. Https://Doi.Org/10.1186/S13023-023-02785-4

[18]. Hett, K., Johnson, H., Coupé, P., Paulsen, J., Long, J., & Oguz, I. (2020). Tensor-Based Grading: A Novel Patch-Based Grading Approach For The Analysis Of Deformation Fields In Huntington's Disease. Neuroimage, 208, 116413. Https://Doi.Org/10.1016/J.Neuroimage.2019.116413

[19]. Greener, J. G., Kandathil, S. M., & Jones, D. T. (2018). Deep Learning Extends De Novo Protein Modelling Coverage Of Genomes Using Iteratively Predicted Structural Constraints. Nature Communications, 9(1), 1–10. Https://Doi.Org/10.1038/S41467-018-06610-0

[20]. Ouwerkerk, J., De Bot, S., Wolstencroft, K., & Mina, E. (2023). Machine Learning In Huntington's Disease: Exploring The Enroll-HD Dataset For Prognosis And Driving Capability Prediction. Orphanet Journal Of Rare Diseases, 18(1), 1–13. Https://Doi.Org/10.1186/S13023-023-02785-4

[21]. Uma Kumari,Gunika Nagpal "NEXT GENRATION SEQUENCE ANALYSIS OF AMYLOID PRECURSOR-LIKE PROTEIN 2 (APLP2) E2 DOMAIN IN ALZHEIMER'S DISEASE", International Journal Of Emerging Technologies And Innovative Research , Vol.12, Issue 1, Page No. Ppd72-D79, 2025,

[22]. Uma Kumari ,Keya Pacholee "In-Silico Drug Discovery-Based Approach To Treat Impairments In Patients Of Alzheimer's Disease", International Journal Of Emerging Technologies And Innovative Research (Www.Jetir.Org), ISSN:2349-5162, Vol.10, Issue 12, Page No.B236-B245, December-2023, Http://Doi.One/10.1729/Journal.37001

[23]. Uma Kumari, Meenakshi Pradhan, Saptarshi Mukherjee, Sreyashi Chakrabarti;, Ngs Analysis Approach For Neurodegenerative Disease With Biopython",2023 ,Volume 10,Issue 9 , Http://Doi.One/10.1729/Journal.36043

[24]. Uma Kumari,Umang Uniyal (First Author)" Next Genration Sequencing Data Analysis Of A25T Transthyretin Structure Associated With Parkinson's Disease", International Journal Of Emerging Technologies And Innovative Research ,Jetir,Vol.12, Issue 3, Page No.B612-B622, March-2025,: Http://Doi.One/10.1729/Journal.44079