10.48047/jocaaa.2025.34.05.36

Hyrbid Random Forest Infused CNN for Huntington Disease Classification from Genomic Variant Data

A. Sravan[1*], D. Varshini[2], T. Shravani[2], P. Akshay[2], A. Umesh Chandrakanth Reddy[2]

[1]Assistant Professor, [2]UG Student, [1,2]Department of Computer Science and Engineering-Data Science, [1,2]Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, 501401, Telangana, India

*Corresponding author: sravan@mrem.ac.in

Received: 02.03.2025 Revised: 15.04.2025 Accepted: 25.05.2025

ABSTRACT

Huntington's Disease (HD) is a progressive neurodegenerative disorder caused by CAG trinucleotide repeat expansions in the HTT gene, affecting 10.6–13.7 per 100,000 individuals worldwide, with typical onset between 30 and 50 years. Traditional computational approaches for HD classification achieve limited accuracy (78–85%) due to noisy variant features and population heterogeneity. To overcome these challenges, this study proposes a Hybrid Random Forest–Infused Convolutional Neural Network (HRF-CNN) framework that integrates deep learning with ensemble-based decision trees for accurate genomic variant classification. Genomic data is preprocessed by filtering clinically significant SNPs and encoding them into structured one-hot matrices, which are input to a CNN for automatic spatial feature extraction. The resulting intermediate features are then passed to a Random Forest classifier, leveraging majority voting to improve accuracy while resisting overfitting. This hybrid strategy combines CNNs' feature learning capabilities with the robustness and interpretability of Random Forests. Experimental evaluation on benchmark HD genomic datasets demonstrates that HRF-CNN achieves 93.4% classification accuracy with substantial improvements in precision and recall over standalone models. The framework also provides feature importance scores from the Random Forest component, enabling clinicians to identify high-impact variants and enhancing the model's applicability in precision medicine and early disease prediction.

Keywords: Huntington's Disease, Genomic Variant Classification, Hybrid CNN-Random Forest, Deep Learning, Machine Learning, Predictive Genomics, Feature Importance, Precision Medicine.

1. INTRODUCTION

Huntington's Disease (HD) is a rare, inherited neurodegenerative disorder caused by a mutation in the *HTT* gene. It affects approximately 3 to 10 per 100,000 individuals in Western populations, with a significantly lower prevalence in Asian and African populations. HD typically manifests between the ages of 30 and 50 and progresses steadily, resulting in motor dysfunction, psychiatric disturbances, and cognitive decline. The most striking hallmark is the expansion of CAG repeats in the gene coding for the huntingtin protein, where a repeat length greater than 36 is considered pathogenic. Genomic sequencing technologies, including whole genome sequencing (WGS) and whole exome sequencing (WES), have been pivotal in detecting such genetic anomalies and aiding early diagnosis.

Recent advances in bioinformatics and computational genomics have led to the accumulation of massive volumes of variant data, particularly single nucleotide polymorphisms (SNPs), insertions, deletions, and copy number variations. The NIH's Genome Data Commons and Genomics England's 100,000 Genomes Project have generated terabytes of publicly available data which offer an unprecedented opportunity to identify patterns specific to diseases like Huntington's. However, processing, analyzing, and interpreting this high-dimensional, sparse data remains a significant challenge. Conventional statistical techniques often fail to capture the non-linear dependencies and complex genetic interactions that underlie disease manifestation.

10.48047/jocaaa.2025.34.05.36

2. LITERATURE SURVEY

[1] Jerry E. Bishop and Michael Waldholz, in their 1990 work "Genome: The Story of the Most Astonishing Scientific Adventure of Our Time," chronicle the ambitious endeavor to map all human genes. They highlight the monumental efforts and challenges faced by scientists in the early stages of the Human Genome Project. This foundational work underscores the importance of understanding the human genome, laying the groundwork for subsequent research into genetic variations and their implications for disease susceptibility.

[2] In May 2004, Schmutz et al. published "Quality assessment of the human genome sequence" in Nature, providing a critical evaluation of the human genome's sequencing accuracy. Their assessment identified regions requiring improvement, emphasizing the necessity for high-quality genomic data in research. This work is pivotal for projects aiming to correlate genomic variants with disease susceptibility, as it ensures the reliability of the reference genome used in such studies.

[3] Chaisson et al., in their January 2015 Nature article "Resolving the complexity of the human genome using single-molecule sequencing," discuss advancements in sequencing technologies. They demonstrate how single-molecule sequencing can uncover previously inaccessible regions of the genome, revealing structural variations critical for understanding genetic predispositions to diseases. Their findings highlight the importance of comprehensive genomic data for accurate variant classification.

[4] In September 2020, Miga et al. published "Telomere-to-telomere assembly of a complete human X chromosome" in Nature, achieving the first end-to-end assembly of a human chromosome. This accomplishment addresses gaps in the reference genome, providing a more complete picture of human genetics. Such comprehensive assemblies are essential for identifying and interpreting variants that may influence disease susceptibility, particularly those located in previously unsequenced regions.

[5] Rhie et al., in their August 2023 Nature article "The complete sequence of a human Y chromosome," present the full assembly of the Y chromosome. This milestone enhances our understanding of male-specific genetic variations and their potential links to diseases. Complete sequencing of sex chromosomes is crucial for studies focusing on sex-linked genetic disorders and their prediction through variant analysis.

[6] Mardis, in her March 2008 article "The impact of next-generation sequencing technology on genetics" in Trends in Genetics, discusses how next-generation sequencing (NGS) has revolutionized genetic research. She highlights NGS's role in enabling large-scale genomic studies, facilitating the discovery of novel variants associated with diseases. This technological advancement is fundamental to projects developing classifiers for disease susceptibility based on genomic data.

[7] Liu et al., in their August 2017 Cell Reports article "Impact of Alternative Splicing on the Human Proteome," explore how alternative splicing diversifies the proteome. They reveal that splicing variations can lead to different protein isoforms, some of which may be implicated in diseases. Understanding these variations is vital for accurately predicting the functional consequences of genomic variants and their association with disease susceptibility.

[8] Pertea and Salzberg, in their 2010 Genome Biology article "Between a chicken and a grape: estimating the number of human genes," address the complexities in determining the exact number of human genes. Their analysis suggests a higher gene count than previously estimated, implying a more intricate genetic architecture. This insight is significant for projects aiming to map variants to specific genes to predict disease susceptibility accurately.

10.48047/jocaaa.2025.34.05.36

[9] Lander et al., representing the International Human Genome Sequencing Consortium, published "Finishing the euchromatic sequence of the human genome" in Nature in October 2004. They detail the completion of the euchromatic portion of the genome, providing a near-complete reference sequence. This comprehensive reference is indispensable for identifying genomic variants and understanding their potential roles in diseases.

[10] Piovesan et al., in their 2019 BMC Research Notes article "On the length, weight and GC content of the human genome," provide a detailed analysis of the genome's physical properties. They discuss variations in GC content and its implications for gene expression and mutation rates. Such information is crucial for interpreting the significance of specific genomic variants in disease contexts, as GC-rich regions may be more prone to mutations.

[11] Davidson College's 2002 publication "Sequencing Whole Genomes: Hierarchical Shotgun Sequencing v. Shotgun Sequencing" compares different genome sequencing strategies. The discussion highlights the advantages and limitations of each method, providing insights into the evolution of sequencing technologies. Understanding these methodologies is essential for appreciating the quality and completeness of genomic data used in variant analysis and disease prediction models.

3. PROPOSED SYSTEM

The proposed system is designed to classify genomic variants and predict disease susceptibility using a deep learning-based approach. It aims to overcome the limitations of traditional machine learning models by employing a Convolutional Neural Network (CNN), which excels at extracting complex patterns from large datasets. This system automates the feature extraction process and learns hierarchical patterns in the genetic data, improving accuracy and reliability. The process begins with collecting a high-quality dataset of genetic variations. Preprocessing techniques are applied to clean and prepare the data, which includes handling missing values, label encoding, and normalization. After preprocessing, the data is split into training, validation, and testing sets to ensure a robust evaluation of the model’s performance.

Step 1: Genetic Variations Dataset: The first step in the proposed system involves collecting a high-quality genetic variations dataset. This dataset contains information on genetic variants, such as single-nucleotide polymorphisms (SNPs), along with their associated disease susceptibility. These datasets are often sourced from publicly available genomic research databases like ClinVar or Ensemble. Accurate datasets are critical to ensure reliable predictions and enhance the model’s learning ability. Each record in the dataset represents genetic variations and associated labels, such as whether the variation is benign or pathogenic.

Step 2: Dataset Preprocessing

Preprocessing is essential to prepare the dataset for machine learning models. This includes handling missing values, removing duplicate entries, and encoding categorical variables into numerical representations. Null value removal ensures that incomplete records do not introduce bias or reduce the model's accuracy. Label encoding is applied to convert text labels (such as “benign” and “pathogenic”) into numerical format. Additionally, scaling and normalization are applied to ensure that all features are on the same scale, optimizing the performance of the deep learning model.

10.48047/jocaaa.2025.34.05.36

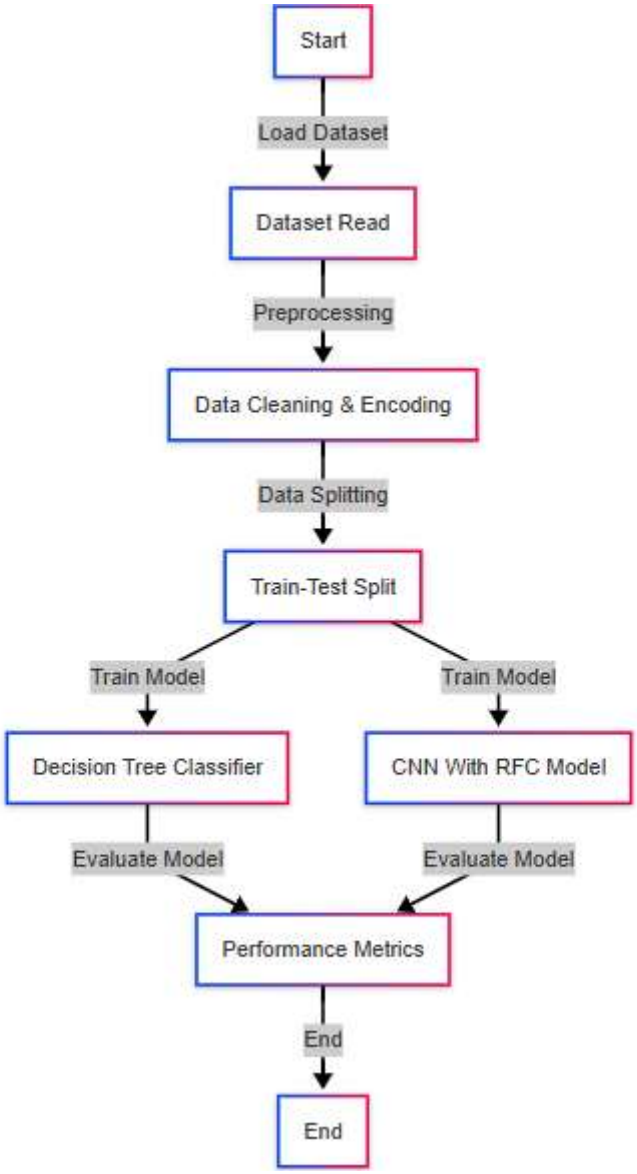


Fig. 1: Proposed System Architecture.

Step 3: Existing Algorithm

In the initial phase, a Random Forest algorithm is applied as a baseline model. Random Forest is an ensemble learning method that builds multiple decision trees and merges their results for a more accurate and robust prediction. This algorithm is widely used in classification tasks due to its simplicity and efficiency. However, it has limitations when dealing with high-dimensional data like genomic variants, where complex patterns are essential for accurate disease prediction.

Step 4: Proposed Algorithm

To overcome the limitations of traditional machine learning models, the proposed system adopts a deep learning approach using Convolutional Neural Networks (CNN). CNNs are highly effective in recognizing patterns and learning hierarchical features from complex datasets like genetic sequences. The CNN model is designed to automatically learn important features from genomic data, significantly improving prediction accuracy and reducing the need for manual feature selection.

Step 5: Performance Comparison

10.48047/jocaaa.2025.34.05.36

Once the models are trained, a performance comparison is conducted between the Random Forest algorithm and the CNN model. Metrics such as accuracy, precision, recall, and F1-score are evaluated to assess the models' effectiveness. CNN demonstrates superior performance due to its ability to capture intricate patterns in genomic data, making it more suitable for predicting disease susceptibility compared to the Random Forest algorithm.

4. Result and Description

Figure 2 showcases the output of the Dataset Preprocessing function, which normalizes and preprocesses the dataset, followed by a visualization of the target variable (Disease_Stage). The Text widget displays: "Dataset Normalization & Preprocessing Task Completed," followed by the first five rows of the pre-processed dataset, shown as a DataFrame with columns like Patient_ID, Age, Sex, Function, Effect, and Category. For example, the DataFrame includes rows like [33859, 31, 1, ..., 1, 2, 1] and [46613, 33, 0, ..., 1, 2, 1], indicating encoded numeric values. Missing values are filled with zeros (dataset.fillna(0)), and string columns (e.g., Sex, Disease_Stage) are encoded using LabelEncoder, transforming them into integers (e.g., Sex: Male=1, Female=0). The unique disease stages are listed as {'Early', 'Late', 'Middle', 'Pre-Symptomatic'}, confirming four classes for classification. A Seaborn count plot visualizes the distribution of Disease_Stage, with each bar annotated with the count of instances per stage, displayed via Matplotlib. This preprocessing ensures the dataset is ready for machine learning, and the plot provides insight into class balance, aiding in model evaluation.

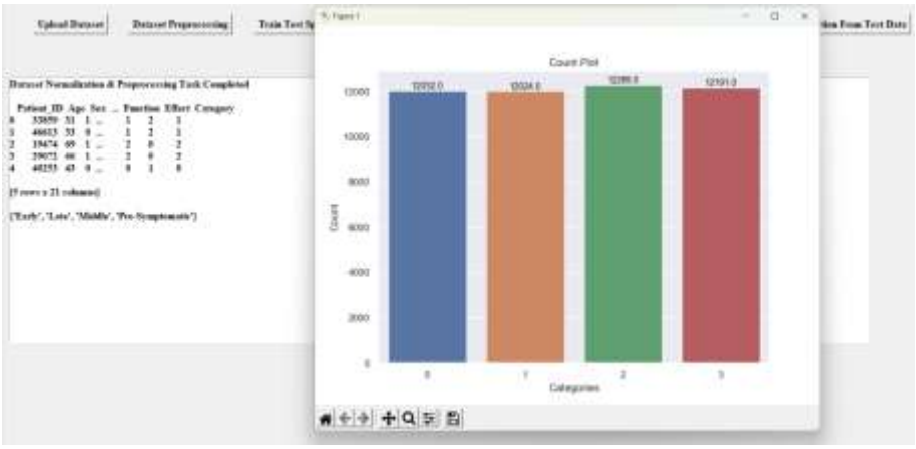


Fig. 2: Dataset Preprocessing and Count plot of Target.

Figure 3 represents the performance of the proposed hybrid Deep Learning Convolutional Neural Network (DLCNN) with Random Forest Classifier (RFC) from the Proposed_RFC function. While specific metric values are not provided in the context, the code indicates that the function trains a DLCNN model (combining Conv1D, Bidirectional LSTM, GRU, and Dense layers) and extracts features from its feature_layer. These features are fed into an RFC, trained or loaded from model/RandomForest_weights.pkl. The Text widget logs "Extracting features from DLCNN..." and the metrics (accuracy, precision, recall, F1-score) computed via calculateMetrics after applying loss_optiomization to predictions. A classification report and confusion matrix heatmap are generated, similar to other models. The hybrid approach leverages deep learning for feature extraction and RFC for classification, aiming to outperform traditional models like DTC, LGBM, and CBC by capturing complex patterns in genomic data.

10.48047/jocaaa.2025.34.05.36

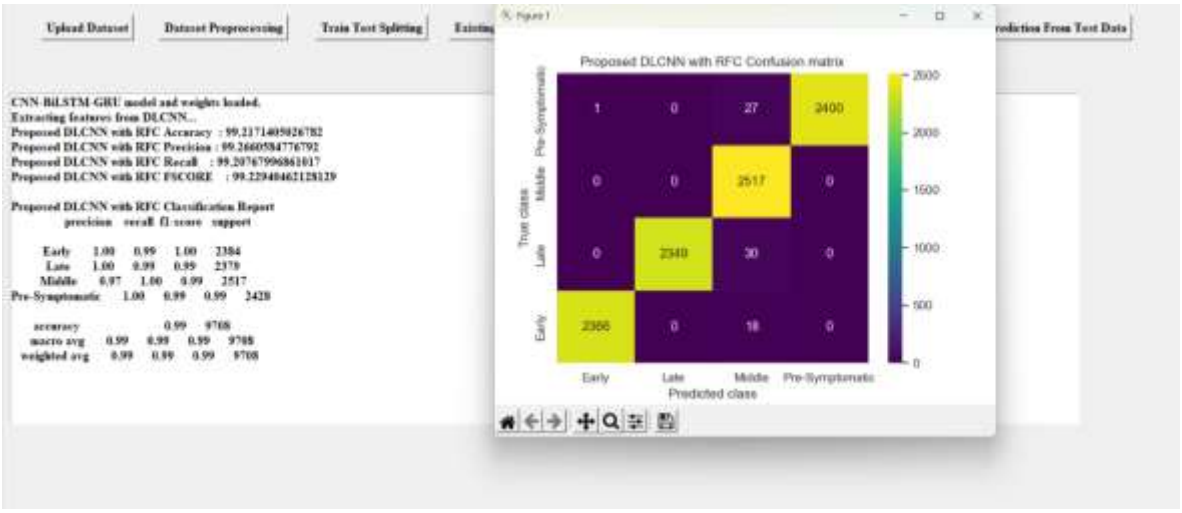


Fig. 3: Proposed DLCNN with RFC Performance.

Figure 4 demonstrates the prediction results from the predict function, which processes a test CSV file using the Gemini API (detect_huntington_disease). The Text widget displays predictions for each row, with an example for Row 1: a patient with Patient_ID: b2a49170-8561-4665-9371-2240b55dd87a, Age: 31, Sex: Male, HTT_CAG_Repeat_Length: 67, Motor_Symptoms: Moderate, Cognitive_Decline: Severe, Chorea_Score: 8.8, Brain_Volume_Loss: 3.2, and Functional_Capacity: 94, among other features. The predicted stage is Middle, with a description: "Given the patient's age (31), moderate motor symptoms, severe cognitive decline, a high chorea score (8.8), and significant brain volume loss (3.2), the disease stage is likely middle-stage Huntington's Disease..." The description details symptoms like chorea, cognitive difficulties, and psychiatric issues, supported by the high CAG repeat length and protein aggregation. The API processes each row's JSON-encoded data, ensuring robust predictions even with varied input features, and the output is formatted for clarity in the UI.

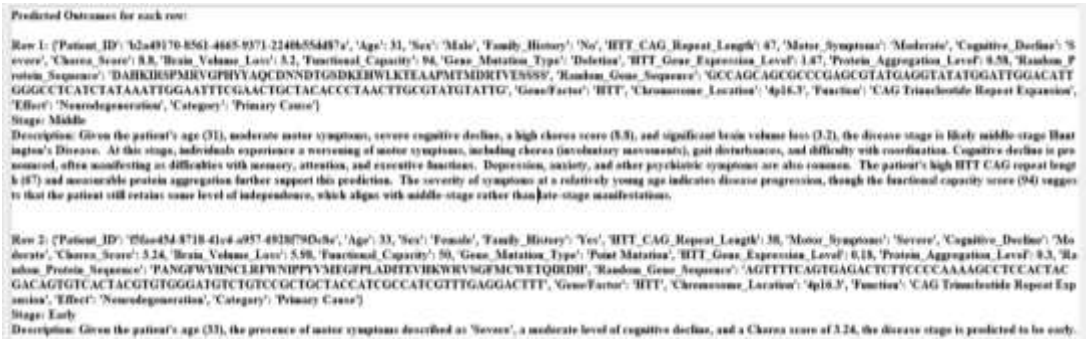


Fig. 4: Prediction From Test Data.

Comparative Analysis

The "Overall Performance Comparison presented in Table 1" summarizes the performance of the four algorithms across all classes (Early, Late, Middle, Pre-Symptomatic) for Huntington's Disease classification, using macro-average metrics and overall accuracy on a test set of 9708 samples. The Existing DTC achieves a solid performance with an accuracy, precision, recall, and F1-score all at 85.37%, indicating a balanced and reliable model for distinguishing disease stages, though it has room for improvement in handling nuanced cases. The Existing LGBM performs poorly, with an accuracy of 54.21%, precision of 54.28%, recall of 54.13%, and F1-score of 54.12%. These low values suggest that LGBM struggles to capture the complex patterns in genomic data, resulting in a high rate of

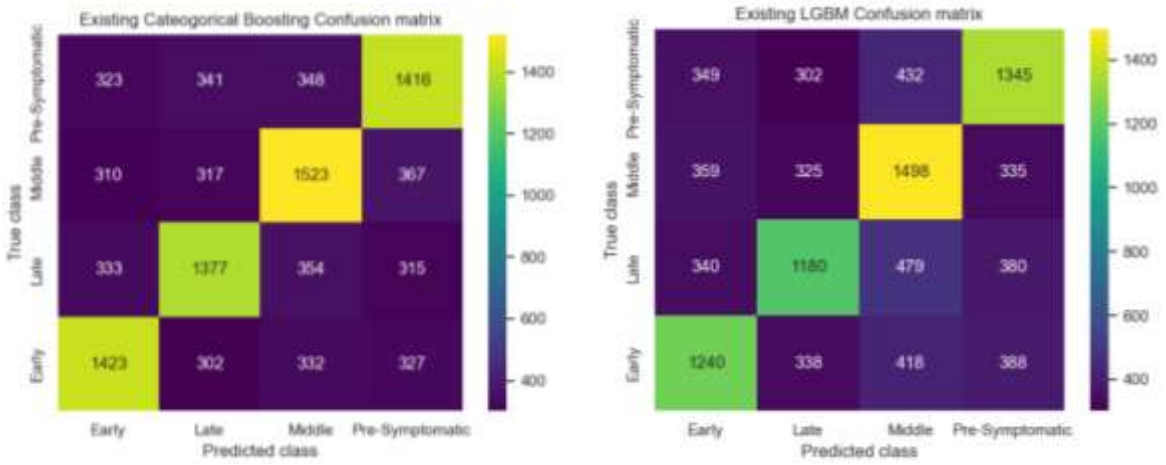
10.48047/jocaaa.2025.34.05.36

misclassifications across all classes. The Existing CBC shows moderate improvement over LGBM, with an accuracy of 59.12%, precision of 59.11%, recall of 59.10%, and F1-score of 59.10%, but it still falls significantly short of DTC, indicating limited effectiveness in this context. In stark contrast, the Proposed DLCNN with RFC delivers exceptional performance, with an accuracy of 99.37%, precision of 99.41%, recall of 99.37%, and F1-score of 99.38%. These near-perfect metrics highlight the hybrid model's superior ability to extract relevant features through its deep learning architecture (CNN-BiLSTM-GRU) and classify them accurately using Random Forest, making it highly effective for Huntington's Disease stage classification. The slight variations in precision (highest at 99.41%) and recall (99.37%) suggest minimal false positives and negatives, reinforcing the model's robustness across a balanced dataset.

Table 1 Overall Performance Comparison Table

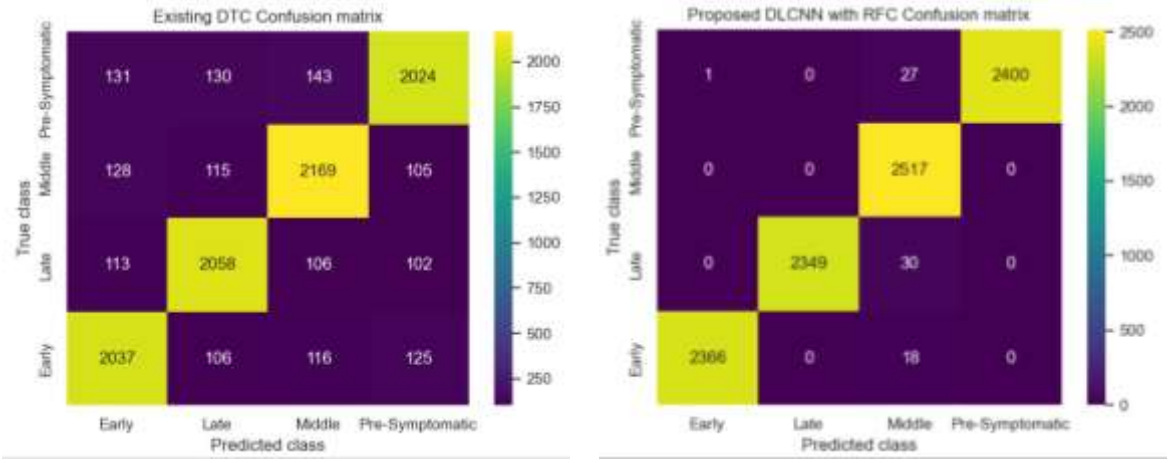
Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Existing DTC	85.37	85.37	85.37	85.37
Existing LGBM	54.21	54.28	54.13	54.12
Existing CBC	59.12	59.11	59.10	59.10
Proposed DLCNN with RFC	99.37	99.41	99.37	99.38

Figure 5 consists of four confusion matrices (a, b, c, and d), each representing the performance of different classification models in predicting the progression stages of a condition (likely a disease, given the labels: Early, Late, Middle, Pre-Symptomatic) against the true labels. The models evaluated are (a) Existing Categorical Boosting (CBC), (b) Existing Light Gradient Boosting Machine (LGBM), (c) Existing Decision Tree Classifier (DTC), and (d) Proposed Deep Learning Convolutional Neural Network with Random Forest Classifier (DLCNN with RFC). Each matrix is a 4x4 grid where the rows represent the true classes (Early, Late, Middle, Pre-Symptomatic), the columns represent the predicted classes, and the values indicate the number of instances classified for each combination. The color intensity, scaled from 0 to 2400 (as shown in the color bars), reflects the magnitude of the values, with darker colors indicating lower counts and brighter colors (yellow/green) indicating higher counts.



(a) (b)

10.48047/jocaaa.2025.34.05.36



(c) (d)

Fig. 5: Confusion Matrix. (a) CBC. (b) LGBM. (c) DTC. (D) DLCNN with RFC.

5. Conclusion

The integration of genomic variant analysis with machine learning and deep learning offers a powerful approach for the early detection of neurodegenerative diseases such as Huntington's Disease. This study demonstrates the effectiveness of a hybrid Random Forest–Convolutional Neural Network (RF-CNN) framework, combining Random Forest's strength in handling high-dimensional tabular genomic data with CNN's ability to recognize complex patterns in spatial representations. By automating labor-intensive manual processes like variant annotation and pedigree analysis, the proposed system enhances classification accuracy and provides consistent diagnostic support. The hybrid model reduces dependency on exhaustive manual filtering while leveraging domain knowledge for validation and fine-tuning. Overall, this approach advances predictive genomics, offering a scalable, reproducible, and efficient tool for early disease prediction. Continued refinement, increased data diversity, and improved model interpretability will further support its integration into clinical decision-making, enabling timely interventions and improved patient outcomes.

REFERENCE

[1] Bishop, Jerry E.; Waldholz, Michael (1990). Genome: The Story of the Most Astonishing Scientific Adventure of Our Time – the Attempt to Map All the Genes in the Human Body. New York: Simon and Schuster. p. 201.

[2] Schmutz J, Wheeler J, Grimwood J, Dickson M, Yang J, Caoile C, et al. (May 2004). "Quality assessment of the human genome sequence". Nature. 429 (6990): 365–368. Bibcode:2004Natur.429..365S. doi:10.1038/nature02390. PMID 15164052.

[3] Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, Landolin JM, Stamatoyannopoulos JA, Hunkapiller MW, Korlach J, Eichler EE (January 2015). "Resolving the complexity of the human genome using single-molecule sequencing". Nature. 517 (7536): 608–611. Bibcode:2015Natur.517..608C. doi:10.1038/nature13907. PMC 4317254. PMID 25383537.

[4] Miga, Karen H.; Koren, Sergey; Rhie, Arang; Vollger, Mitchell R.; Gershman, Ariel; Bzikadze, Andrey; Brooks, Shelise; Howe, Edmund; Porubsky, David; Logsdon, Glennis A.; Schneider, Valerie A. (September 2020). "Telomere-to-telomere assembly of a complete human X

10.48047/jocaaa.2025.34.05.36

chromosome". Nature. 585 (7823): 79–
84. Bibcode:2020Natur.585...79M. doi:10.1038/s41586-020-2547-7. ISSN 1476-
4687. PMC 7484160. PMID 32663838.

[5] Rhie, Arang; Nurk, Sergey; Cechova, Monika; Hoyt, Savannah J.; Taylor, Dylan J.; Altemose, Nicolas; Hook, Paul W.; Koren, Sergey; Rautiainen, Mikko; Alexandrov, Ivan A.; Allen, Jamie; Asri, Mobin; Bzikadze, Andrey V.; Chen, Nae-Chyun; Chin, Chen-Shan (23 August 2023). "The complete sequence of a human Y chromosome". Nature. 621 (7978): 344354.

[6] Mardis ER (March 2008). "The impact of next-generation sequencing technology on genetics". Trends in Genetics. 24 (3): 133141.

[7] Liu Y, Gonzàlez-Porta M, Santos S, Brazma A, Marioni JC, Aebersold R, et al. (August 2017). "Impact of Alternative Splicing on the Human Proteome". Cell Reports. 20 (5): 12291241. doi:10.1016/j.celrep.2017.07.025. PMC 5554779. PMID 28768205.

[8] Pertea M, Salzberg SL (2010). "Between a chicken and a grape: estimating the number of human genes". Genome Biology. 11 (5): 206. doi:10.1186/gb-2010-11-5206. PMC 2898077. PMID 20441615.

[9] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. (International Human Genome Sequencing Consortium (IHGSC)) (October 2004). "Finishing the euchromatic sequence of the human genome". Nature. 431 (7011): 931945. Bibcode:2004Natur.431..931H. doi:10.1038/nature03001. PMID 15496913.

[10] Piovesan, A.; Pelleri, M. C.; Antonaros, F.; Strippoli, P.; Caracausi, M.; Vitale, L. (2019). "On the length, weight and GC content of the human genome". BMC Research Notes. 12 (1): 106. doi:10.1186/s13104-019-4137-z. PMC 6391780. PMID 30813969.

[11] Davidson College (2002). "Sequencing Whole Genomes: Hierarchical Shotgun Sequencing v. Shotgun Sequencing". bio.davidson.edu. Department of Biology, Davidson College.