INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 3, MARCH 2026 | ISSN: 2456-4184 | IJNRD.ORG

Predictive Modeling Of Repetitive DNA Elements For Identification And Classification Using Machine Learning

[1]B. Ramji, [2]Bandapalli Prakash, [3]Donthula Abhinav, [4]Parupally Varshitha
[1]Assistant Professor, [2]Student, [3]Student, [4]Student
[1234]Department of Computer Science and Engineering (Data Science)
[1234]CMR Technical Campus Hyderabad, India

Abstract : Human DNA is very complex and consists of tandem and non- tandem repeats. The expansion of these tandem repeats is linked to several neurological and developmental disorders. Existing tools struggle to identify these patterns, and they also do not provide the diseases linked to that raw DNA sequence. Therefore, we developed a computational tool using machine learning algorithms, specifically Logistic Regression, Random Forest, and XGBoost to predict whether a given DNA sequence is a tandem repeat or a non- tandem repeat. Additionally, the system identifies the gene linked to that sequence along with diseases associated with that particular gene. This framework serves as a valuable analytical resource for clinicians and research centres to process raw DNA sequences.

Keywords - Human DNA, Tandem Repeats, Non-Tandem Repeats, Repeat Expansion, Machine Learning, Logistic Regression, Random Forest, XGBoost, DNA Sequence Classification, Gene Identification, Disease Associated

I. INTRODUCTION

Repetitive DNA sequences are abandoned in eukaryotic cells which can affect gene functions contributing to diseases. These Repetitive sequences include tandem repeats and transportable elements by constantly changing and moving around. The biggest problem is when these stack of repeat sequences get unnaturally huge, it causes major diseases like Huntington's and Fragile X. These repeats actively mess with how genes turn off causing our genome chromatin unstable or chaotic.

Over 20 years different tools and methods have been tried to fix this issue. TRF(Tandem Repeats Finder) was the first method that gave people a way to start solving this problem. Later skipping important steps techniques are such as alignment methods and mismatch techniques were used to fill this gap in the process. However, these methods were not impact, due to several limitations. New variations in DNA could bypass these checks and complex pattern, causing difficulties in computer programming and scanning pull DNA sequences because of very slow and computational power was limited. Because of this, the process often becomes slow and exhausting at the same time challenging.

New technologies improve the process. Instead of using old methods Todd's team introduced a new methods like HMMSTR and Mousavi's. These methods help scientists to identify repetitive DNA patterns more accurately. Although these methods shows strong improvement, their performance still depends on long DNA sequences and the amount of data. In small laboratories these creates difficulties with limited resources. However more studies focused mainly on identifying and classifying these elements rather than predicting their role in disease development.

AI was introduced into this field for research. Zhang's team demonstrated that deep learning has identified and labelled, how genes could completely change. Liao's group detected repetitive DNA sequences using ML algorithms, which was considered a bold approach at that time. Fotsing's team found how these repeated tandems are linked to change in gene expression. Although many of these tools reports strong performance, but understanding their result is still difficult and turning them into real medical applications is still hard.

To address this research gaps we need to develop a analytical framework. This system extract features such as pattern frequency, GC content and the occurrence rate of the repeat, and inputs them into a ML model. This model classifies Repeated DNA sequences which are likely or unlikely to be associated with repeated related disorders. In addition a manual curated gene disease reference database was included to ensure that the predictions are supported by real biological evidence.

The integration of machine learning and deep learning methods with genomic features result in fast and easy interruption with quality output. This approach improves the detection of tandem repeats with their associated disease risk while efficiently handling large scale genomic data.

1.1 Key Contribution

a. Feature Extraction:

We have implemented a custom feature extraction method that measures the maximum unbroken chain length of DNA motifs. This allows the ML model to understand actual biological disease thresholds instead of just counting random pattern occurrences.

b. Handling Imbalanced Genomic Data:

Since we have too few disease samples compared to healthy ones, to solve this we used Random Oversampling to balance our training dataset. By training an XGBoost classifier on this balanced data, we achieved a highly precise model that eliminates false positive diagnoses

c. Integrated Clinical Mapping:
We have created a manually verified gene–disease mapping database. When the ML model detects a gene, it maps the gene to the database and fetches the associated diseases.

II. LITERATURE REVIEW

Repetitive DNA sequences have been an important focus in genomics for many years. These tandem repeats play a major role in evolution, genome organization, and the development of several diseases. Many researchers have developed both computational and laboratory methods to detect and analyse these repeats. However, important gaps still remain, especially in prediction accuracy and model stability.

Todd et al. [1] developed a hybrid model that combines tandem repeat detection with a Hidden Markov Model (HMM) to identify disease-related repeats. This method achieved high accuracy in detecting large and complex repeat regions. However, it required deep sequencing and high computational resources, which limited its practical use. In contrast, the proposed system reduces both cost and workload by using machine learning to predict and extract DNA sequences without relying on deep sequencing.

Zhang et al. [2] reviewed traditional and deep learning methods used to detect genes and regulatory regions. Their study mainly focused on gene functionality rather than predicting diseases related to tandem repeats. The work aimed at combining statistical sequence analysis with predictive models to identify high-risk genomic regions.

Liao et al. [3] used statistical and alignment-based techniques to detect tandem repeats in the human genome. Although their method performed well on known reference sequences, it showed weak performance on unknown regions. This limitation was later addressed by using machine learning to learn repeat patterns directly, enabling better generalization and detection of previously unknown repeats.

Duitama et al. [6] studied large-scale tandem repeat variations across human genomes. Their research explained how repeat regions change and diversify over time, providing valuable biological insights. However, their work was descriptive and did NOT attempt to predict disease risk. The proposed model goes a step further by using repeat-related features to predict the likelihood of disease, not just describe repeat patterns.

In summary, previous research has significantly improved our understanding of repetitive DNA and its distribution in the genome. However, most studies have not focused enough on predicting disease risk from tandem repeats in a scalable and interpretable way. The machine learning model proposed in this study addresses this limitation by effectively integrating biological knowledge with computational techniques, enabling learning, adaptability, and accurate disease risk prediction rather than simple pattern description.

III. PROPOSED METHODOLOGY

Our framework uses three crucial steps to build this tool. Step 1 is data acquisition, where we collect the data from different websites and use different filtration processes to make it ready for machine learning training. Step 2 is the machine learning pipeline, where we use different algorithms and features to train the model with the data that we have processed. Step 3 is visualization, basically it acts as a bridge to connect the different sources we have, and it presents the user with an interactive user interface where the user can simply enter a raw DNA sequence and get a report.

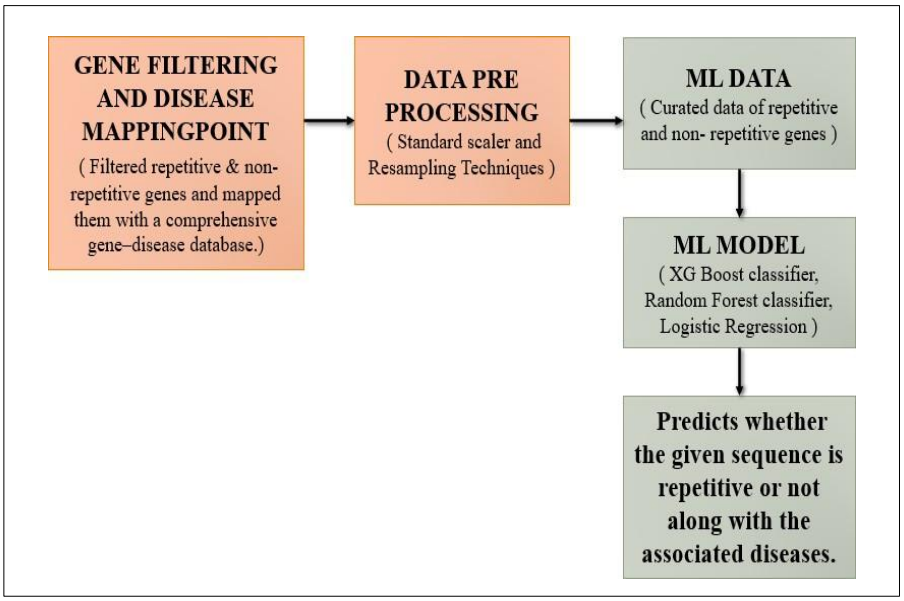

Fig 1: Architecture

3.1 Data Acquisition

The first step is data acquisition, which involves collecting the data. We collected human genomic data from the NCBI website; this data is very large and raw. Since the data is raw, we performed filtration techniques and feature extraction to make it usable for machine learning training. First, we filtered different motifs along with their frequency count as well as their starting index. A motif is basically a small part of the genome, for example, "ATG." The frequency indicates how many times a motif is been repeated in entire DNA sequence and starting index signifies the start of the motif. This index is useful for finding the gene it has been associated to, which we will discuss in further.

Once we obtained the starting indices of all the repetitive motifs, we queried the NCBI database. By using these indices, we retrieved the gene name, and by knowing the gene name, we identified the diseases associated with it. We created a local database which is called the "Gene Database," where we manually mapped the genes collected from the NCBI website. In this process, we mapped all the genes along with their associated diseases making this a seperate database which we will use in gene lookup to identify the diseases.

Finally, coming to feature extraction: Currently, we use frequency and starting index, and we have also extracted a few more features to make it easier for the machine learning model to predict whether a sequence is a tandem repeat or not. We extracted the percentage of the tandem repeat in the entire DNA sequence, the length of the repeat, entropy and period size.

3.2 Machine Learning Pipeline

Coming to the machine learning pipeline, we used three different algorithms: XGBoost, Logistic Regression, and Random Forest. The reason we used Logistic Regression is to predict whether the given sequence is a tandem repeat or not basically 0 or 1. To do that, the model needs to have two different types of data. Initially, we had only tandem repeats, so we needed to add non-tandem repeats as well to make it easier for the model to predict whether the sequence is a tandem repeat or not. Therefore, we added the exact number of non-tandem repeats to balance the two different types of data.

The reason we used Random Forest is because we have many features. Random Forest helps the machine learning model to pick the best features so that it can learn more from them; essentially, it picks the best features where it learns more, and it increases the efficiency of the model.

Coming to XGBoost, since we have a smaller amount of data even though DNA is very vast, we know that we have limited tandem repeats and also we have an exact count of non-tandem repeats as well, which makes the data very small so in order to increase the efficiency of the model, we used XGBoost for this.

3.3 Tools Implementation

So, we have two different components now: the first one is the Gene Database and the other one is our machine learning model. The Gene Database consists of genes with the respective diseases associated with them. Coming to the machine learning model, this is executed and saved under an extension called .joblib, where this extension can be used in different applications easily. So, this tool basically connects these two—the database and this machine learning model—and makes a visual interface (UI) for the users to enter the raw DNA sequence. Through the machine learning model, it predicts the gene, and then with the help of the Gene Database, our algorithm looks in the database to predict the associated diseases with the gene. Finally, we get whether the sequence is a tandem repeat or not, along with the gene and its associated diseases as well.

IV. Result And Discussion

4.1 Data Flow

The figure 2 visualizes the data flow where it starts from loading the ML models, which are saved as .joblib, and datasets such as the Gene Database. Then, the user is provided with the UI where we have an input for the raw DNA sequence. So, when a user pastes the raw DNA sequence and clicks on predict, our loaded ML model predicts the sequence using Logistic Regression, Random Forest, and XGBoost and classifies it as high risk or low risk; basically, high risk means a tandem repeat and low risk means a non-tandem repeat. After predicting, if it is a high risk, our ML model also gives the gene it is associated with. After getting the gene, we will look up in the Gene Database to get the linked diseases. If we do not get a high risk, we will just display the result without the gene and its diseases.

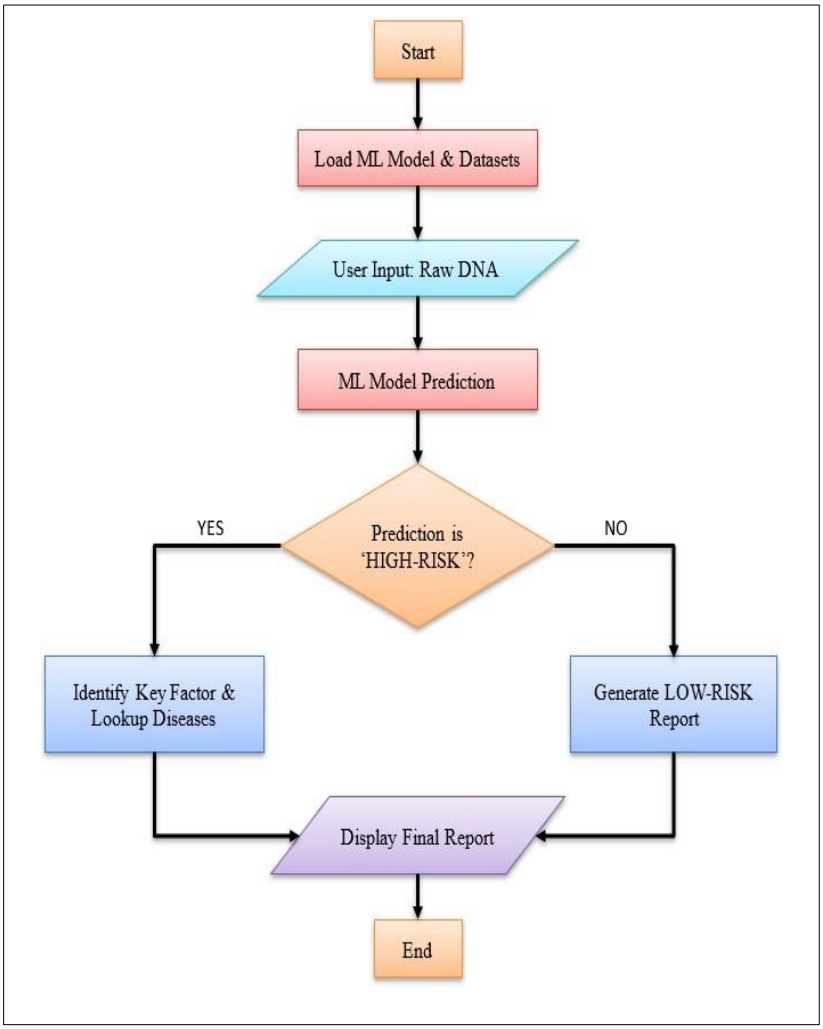

Fig 2: Data flow of the system

4.2 Datasets

Coming to the datasets, first, as we mentioned before, we collected the human DNA file from NCBI which is very large in size. This file is in raw format and needs processing and feature extraction. We used string matching and filtering techniques to extract features like Motif, Frequency count, Starting index, Percentage, Length of the motif, Entropy, and Period size.

We have also created a Gene Database using the starting index and mapped genes with their respective diseases. We created this database manually by looking up the index in NCBI, getting the gene, and finding out its diseases. This database is useful for disease lookup; when the ML predicts a gene, we can get its associated diseases through this database.

4.3 Results

Table 1: Comparison table for previous mode and the current proposed model.

S. NO	MODEL	ACCURACY (%)
1	BLAST	81.2%
2	RepeatMasker	85.6%
3	Proposed (XGBoost)	88.5%

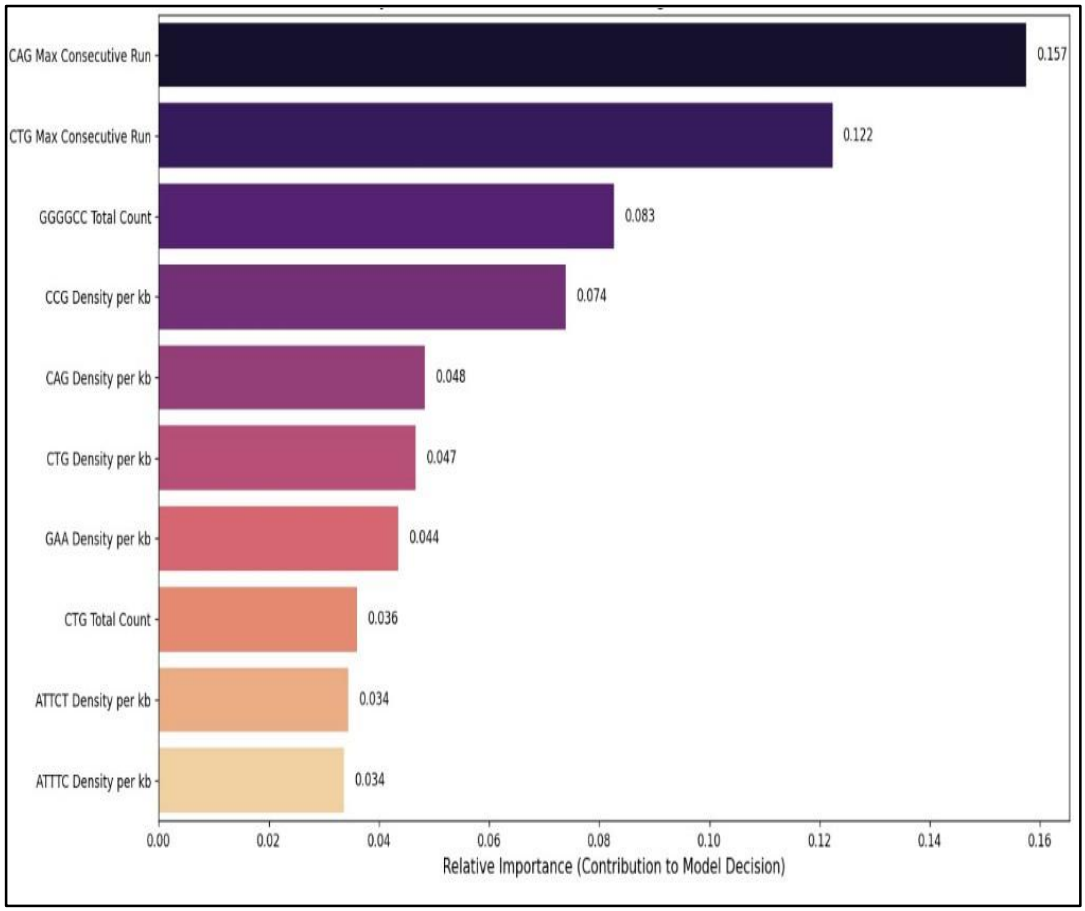

Fig 3: Top 10 genomic features influencing model decision

Table 1 and Figure 3 summarize the performance and internal decision-making process of our proposed machine learning framework.

As shown in Table 1, the XGBoost classifier achieved an overall accuracy of 88.5%. However, evaluating false diagnoses is crucial for clinical tools. In this regard, our model achieved a perfect precision score of 100% (1.00) for identifying pathogenic repeat expansion genes, meaning our model produced zero false-positive errors during testing. Due to the limited availability of rare repetitive disease samples in the training dataset, due to this we got a Recall score of 40% (0.40). This data scarcity limitation will be addressed in future iterations.

To understand how the model achieved such high precision, we analyzed its mathematical decision-making process, illustrated in Figure 3. The chart reveals that the algorithm heavily relied on our engineered features

- specifically the 'Maximum Consecutive Run Length' of CAG and CTG motifs over raw motif frequencies. This proves that our model autonomously learned the true biological mechanism of repeat expansion diseases, validating our feature engineering approach.

V. CONCLUSION

This research develops a machine learning model to predict genomic diseases associated with repetitive DNA sequences. Instead of analysing each DNA sequence manually, this model automatically extracts the key features such as repeat frequency, GC content and genomic positions of repeat clusters. These features are provided to train the model to identify regions with high risk of repeat expansion. It integrates gene disease mapping tool is used to interpret the predictions by linking competition results to real clinical cases there by making outcomes meaningful in biology and medicine.

This experiment proves that XGBoost method approach outperforms traditional genomic tools such as BLAST, RepeatMaster and HMMSTR. The difference in performance, highlighting the ability of flexible machine learning system can manage a very large DNA data set more efficiently. This shows a shift from using rigid rule based software to intelligent AI models that get improve through learning from the data and experience.

Because rare genetic diseases have limited sample sizes, the model's current recall sensitivity (40%) is restricted by the small number of pathogenic cases in the training data. Therefore, future work will primarily focus on expanding the clinical dataset with a higher volume of known pathogenic sequences to improve the recall sensitivity.

VI. REFERENCES

[1] Peter K. Todd, Enhanced detection and genotyping of disease-associated tandem repeats using HMMSTR and targeted long-read sequencing. Nucleic Acids Res 53, 5–15 (2025).
[2] Xingtan Zhang, From tradition to innovation: conventional and deep learning frameworks in genome annotation. Brief Bioinformatics 25, 1–15 (2024)
[3] Xingyu Liao, Repetitive DNA seq detection and its role in the human genome. Commun Biology 6, 954 (2023).
[4] Kidwell. Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A 94, 7704–7711 (1997).
[5] Benson. Tandem repeats finder: program toanalyze DNA sequences. Nucleic Acids Res 27, 573–580 (1999).
[6] Duitama. Large-scale analysis of tandem repeat variability in the human genome. Nucleic Acids Res 42, 5728–5741 (2014).
[7] Louzada. Decoding the Rolep of Satellite DNA in Genome Architecture and Plasticity—An Evolutionary and Clinical Affair. Genes 11, 72 (2020).
[8] Brown. Structure-forming repeats and their impact on genome stability. Curr Opin Genet Dev 67, 41–51 (2021).
[9] Padeken. The epigenetic control of transposable elements in development and in diseases. Front Genet 14, 1282449 (2023).
[10] Cross. The genomic study of repetitive elements in Solea senegalensis reveals multiple impacts of transposable elements in the evolution and architecture of Pleuronectiformes chromosomes. Front Mar Sci 11, 1359531 (2024).
[11] Sotero-Caio. Vertebrate transposable elements: mechanisms, evolution and roles. Mob DNA 8, 3 (2017).
[12] La Spada. Repeat expansion disease: progress and puzzles in disease pathogenesis. Nat Rev Genet 11, 247–258 (2010).
[13] Nelson. The unstable repeats: Three evolving faces of neurological disease. Neuron 77, 825–843 (2013).
[14] Nurk. The complete sequence of a human genome. Science 376, 44–53 (2022).
[15] Hannan. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. Trends Genet 26, 59–65 (2010).
[16] Hannan. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet 19, 286–298 (2018).
[17] Chalopin. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol 7, 567–580 (2015).
[18] Fotsing. The impact of short tandem repeat variation on gene expression. Nat Genet 51, 1652–1659 (2019).
[19] Mousavi. Profiling genome-wide landscape of tandem repeat expansions. Nucleic Acids Res 47, e90 (2019).
[20] Shao. Evolution and diversity of transposable elements in fish genomes. Sci Rep 9, 1–8 (2019).