

GC Skew and Codon Bias Analysis of Human Mitochondrial DNA for Disease Marker Discovery

K.Malarkodi
Assistant Professor
Saveetha School of Engineering
SIMATS, Saveetha University Chennai
malarkodiharini@gmail.com

Abstract— The essential role of mtDNA in cellular energy metabolism creates numerous genetic and metabolic disorders because it mutates. A systematic technique for disease marker identification uses GC skew and codon bias pattern analysis of human mitochondrial DNA sequences according to this research. Scientists applied static window sizes to analyze mtDNA as they processed both GC skew and GC content measurements within each segment. The extracted sequence data from the program underwent normalization processing that created heatmaps which were reduced to 66-dimensional vectors. The vectors included 64 codon frequency measurements together with two compositional statistics. The Extreme Gradient Boosting(XGBoost) classifier obtained its segment training labels from MITOMAP disease associations. The analysis included two features known by researchers as A+T skew and RSCU that boosted their performance as enrichment tools. The SHAP analysis method performed the interpretation of the models. The results show good predictive performance along with a constant feature influence which demonstrates the effectiveness of integrating codon-originated features with combination-based features for finding mitochondrial disease biomarkers. This approach establishes a system that extends and explains as a method for building genetic diagnostic tools.

Keywords—Mitochondrial DNA, GC skew, Codon bias, MITOMAP, XGBoost.

I. INTRODUCTION

Mitochondrial DNA acts as a vital faculty in cellular energy production because it contains genetic material necessary for proper mitochondrial function. The maternal inheritance of mtDNA creates conditions for the formation of mitochondrial diseases in tissues which require significant energy. The oxidative phosphorylation gene defects cause two disorders Leber's hereditary optic neuropathy and MELAS syndrome. Giving precise diagnoses and proper treatments depends heavily on the identification of disease-causing mutations in mtDNA. The compact mitochondrial DNA genome consists of only 16,569 base pairs yet carries 13 protein encoding sequences and 22 transfer RNA sequences and 2 ribosomal RNA sequences as well [1]. Faulty mutations occurring within these regions create problems that result in several different disorders. Diagnosing mitochondrial pathologies remains difficult because mtDNA exhibits high mutational frequency as well as distinct genetic transmission patterns [2]. Advanced detection approaches are required to detect mutations together with functional areas because of these difficult conditions [3]. The analysis of mtDNA benefits from two genomic elements which include GC skew and codon bias studies. The GC skew value shows how G and C nucleotide

frequencies compare with one another and helps scientists understand DNA replication together with transcription patterns. The pattern through which synonymous codons are used non-randomly affects how proteins get synthesized and how genes function. The features affect mitochondrial gene operation and they might correspond with disease-related mutations [4].

The mitochondrial genomic data reveals replication origins through GC skew analysis and it simultaneously indicates locations of functional genomic regions. Research shows that variations in GC skew occur near genes which cause mitochondrial diseases. The analysis of GC skew patterns in mtDNA DNA enables scientists to uncover disease-related mutation indicators which improves mitochondrial gene regulatory comprehension [5]. Among coding sequences in mtDNA there exists substantial non-randomness in codon usage because specific codons are preferred over alternative options. The tRNA availability together with translation efficiency determines this bias. The use of certain limited tRNA molecules in mtDNA creates conditions where specific codons influence protein synthesis performance. The analysis of codon preferences in mitochondrion DNA enables scientists to understand biological limitations through detection of potential disease indicators. The Electron Transport Chain genes MT-ND1 MT-CO1 MT-CYB together with MT-ND1 directly link to diseases [6]. Genes that exhibit mutations lead to mitochondrial impairment which causes the development of neurodegenerative diseases alongside myopathy. The research of disease markers depends strongly on discovering genetic variations specifically in GC skew parameters and codon bias patterns [7].

A high mutation rate together with heteroplasmy makes it challenging to identify disease markers in mtDNA. The study of genomic features including both GC skew and codon bias remains hidden when researchers use traditional mutation-based approaches [8]. These characteristics present a possible solution to identify disease-associated regions more thoroughly than mutation-based detection methods despite their potential weaknesses. Complex genomic data gets evaluated through effective solutions which result from modern computational biology and machine learning capabilities. Through the usage of GC skew and engineered metrics in combination with codon usage machine learning algorithms achieve accurate classification of disease regions in mtDNA sequences. The method provides extensive recognition of possible disease indicators [9].

This research investigates mitochondrial DNA disease-associated mutations together with GC skew and codon bias to determine their interconnections. Advanced classification methods including XGBoost will combine with these features in order to build an identification model for disease markers. This work enhances knowledge about mitochondrial diseases through its development of an early detection computational system.

II. LITERATURE REVIEW

The authors [10] researched the relationships between codon usage bias (CUB) and gene length in 60 neurodegeneration-related genes. The analysis focused on compositional features and relative synonymous codon usage (RSCU) and various nucleotide skews while supporting that gene length demonstrates positive correlation with CUB at both short and long ranges of transcript length (less than 1,200 bp and greater than 2,400 bp). Thus it became evident that GC-terminating codons are preferred while specific codons such as TTA, GTT, GGA link to increased gene size. The research demonstrates that both gene size structure together with biochemical preference patterns shape expression capability while possibly affecting the evolutionary process of neurodegeneration-linked genes.

Researchers at [11] carried out sequencing and comprehensive analysis of mitochondrial genomes extracted from two flea taxa *Paradoxopsyllus custodis* and *Stenischia montanis yunlongensis* taken from Yunnan China plague areas. Each genome measured at 15,375 bp and 15,651 bp in length with the standard animal mitochondrial genes of 37 features. The mitochondrial genomes showed abundant AT content and every protein-coding gene started with an ATN combination. Phylogenetic analyses that used maximum likelihood and Bayesian inference evidence showed Leptopsyllidae exists as paraphyletic taxonomic group. The research brings improved knowledge of flea classification along with essential genetic information that helps advance vector biology research within Siphonaptera.

The researchers at [12] analyzed the codon usage patterns along with codon pair frequencies within 18 depression-related genes stored in the NCBI Genetic Testing Registry. The researchers detected a strong preference for GC-ending codons together with substantial codon usage bias which concentrated mainly on CTG and GTG codons. The research identified the connection between depressive genes through pathways leading to serotonin signaling pathways and dopamine pathways and drug metabolism functions. The RSCU analysis confirmed important differences emerged between housekeeping genes and the studied depression-related genes through their choice of codons. The results showed that the CUB measured statistically significant correlations with two factors: gene measurement and protein measurement. Additionally the skew analyses demonstrated that pyrimidine, amino, and keto elements influenced the system. The data indicates that codon preferences together with nucleotide compositions act as determinants for depression-related gene expression and regulation.

The researchers from [13] derived the complete mitochondrial genome sequence for *Neolamarckia cadamba* which stands as an important tropical tree with medicinal and commercial value. The PacBio sequencing method resulted in genome assembly of two circular molecules that measured 109,836 bp and 305,144 bp while containing 83 genes particularly 40 protein-coding genes (PCGs). The phylogenetic analysis based on 24 PCGs and rps3 sequence showed that *Neolamarckia cadamba* belongs to Gentianales order and the Cinchonoideae subfamily within the family of Rubiaceae. The genome showed an AT-leaning makeup because of its composition that reached 54.82% (A+T). The research represents the initial mitochondrial genome sequence of *N. cadamba* which creates important resources for molecular marker development and genetic diversity investigations.

It [14] examined codon usage bias (CUB) patterns in 263 coding sequences which spanned through 44 human blood group systems in order to establish the genetic structure and detect relationships with neurodegenerative disorders. Blood group genes received classification into two distinct groups after assessing their reported neurodegeneration relationships. RSCU analysis together with GC-rich composition of the genes indicated directional selection because G/C-ending codons were observed strongly preferring leucine-coding CTG codon. The examination of serine and proline codons in the two groups produced different distribution patterns. The occurrence of CpC and GpG dinucleotides matched patterns typically found in neurodegenerative conditions connected to non-CpG methylation. Observational and statistical tests demonstrated distinct codon preferences among genes from both groups that highlighted how natural selection affects gene expression in diseased patients.

Research teams have examined codon usage bias within depression-associated genes and neurodegeneration-related transcripts as well as mitochondrial genomes of plants and insects and blood group genes affecting disease risk. Multiple investigations demonstrate that nucleotide sequences together with preferred codons along with genes of different lengths and natural selection pressures will influence both molecular evolution and gene expression. The valuable findings from previous research have not yet explored the complete genetic processes along with codon usage signatures which we specifically examine in our study. This research builds upon existing knowledge by examining [the codon usage patterns coupled with disease associations in their linkage across multiple biological systems] which leads to an advanced understanding of human health and disease through codon bias.

III. MATERIALS

The study relied on GRCh38 human genome reference assembly with its mitochondrial DNA sequence presented as *Homo_sapiens.GRCh38.dna.chromosome.MT.fa* from the dataset of Kaggle [15]. The Genome Reference Consortium released GRCh38 as an improved human genome reference which increases genomic research accuracy through its advanced quality standards. The DNA sequence of chrMT was specifically used because researchers frequently utilize

it for maternal inheritance investigations and population genetics work and studies of mitochondrial disorders. The widely used and trusted reference provided dependable conditions for the scientific validity of our genomic measurements.

IV. METHODOLOGY

The research investigates disease-related markers in human mitochondrial genomes by analyzing GC skew patterns as well as studying codon usage bias patterns. We used bioinformatics pipelines as data extraction tools to obtain sequence features simultaneously with machine learning methods that discriminated regions between disease-relevant or disease-irrelevant sequences.

A. Data Preprocessing and Sequence Segmentation

At the first step the raw mitochondrial DNA sequence required formatting error cleaning through removal of newlines and excess spaces. After cleaning the sequence it was divided into segments that maintained optional overlap between each other. The length of each segment measured either 300 or 500 base pairs (bp) and shared 50% overlap to provide neighboring segments with identical base sequences.

The segmentation technique enables researchers to study small genomic areas thus increasing precision along with prediction accuracy. The sequence S gets divided mathematically into N segments which can be either overlapping or non-overlapping $\{S_1, S_2, \dots, S_N\}$, through this process. The segments S_i of the sequence S have lengths L_t and $L \in \{300, 500\}$ bp.

$$S_i = S[\text{start}_i : \text{start}_i + L] \quad (1)$$

Where:

- start_i is the starting position of the segment S_i .

The GC skew and content calculation has received pre-processed data transmission.

B. GC Skew and Content Calculation

The features of GC skew and GC content provide valuable information about nucleotide patterns because they relate to mitochondrial genome functionality. The two metrics function together to discover changes in DNA sequence structures and functions which could possibly result in diseases.

GC Skew: The GC skew operates as a metric which shows the unbalance between guanine (G) and cytosine (C) occurrences in DNA sequences. The GC skew computation applies to individual sections within DNA sequences. Well-balanced sequences should display a GC skew near zero but any deviation implies structural or functional important regions. The formula for GC skew is:

$$\text{GC Skew} = \frac{G - C}{G + C} \quad (2)$$

Where:

- G = Number of guanine bases in the segment

- C = Number of cytosine bases in the segment
- The ratio detects genomic areas that show irregularities in DNA replication or transcription patterns.

GC Content: Genomic sequences contain information about DNA composition through the count of both guanine (G) and cytosine (C) nucleotides which calculate GC content. The presence of high GC content suggests functional parts of the genomic sequence including coding regions plus stability-dependent genes. The formula for GC content is:

$$\text{GC Content} = \frac{G + C}{A + T + G + C} \times 100 \quad (3)$$

Where:

- A = Number of adenine bases in the segment
- T = Number of thymine bases in the segment
- G and C are as defined above.

The calculations serve every segment for essential data features that help distinguish disease-linked from non-disease-linked mitochondrial sequences.

C. Codon Usage Analysis

The utilization of codons proves essential because living organisms plus various genetic parts across their genomes show varied frequencies of three-nucleotide sequences. The bias shows where genes are highly active and which protein synthesis modifications may occur.

Codon Frequencies: A 300-500 bp segment undergoes a translation process to generate codons from its DNA nucleotide sequences through triplet nucleotide combinations. A count of these existing codons takes place for each analyzed segment throughout the study. Researchers count the codon frequencies within the DNA segment and normalize these values through total codon division to obtain results. Mathematically, the frequency f_i of codon i in a given segment SSS is computed as:

$$f_i = \frac{\text{Count of Codon}_i}{\text{Total Codons in Segment}} \quad (4)$$

Where:

- f_i = Frequency of codon i in segment S ,
- Count of Codon_i = Occurrences of codon i in segment S ,
- $\text{Total Codons in Segment}$ = Total number of codons in the segment.

The analysis gives complete information about which nucleotides contribute to coding capacity across each segment to expose biologically significant relationships between gene expression and mutations.

D. Heatmap Construction and Feature Vector Flattening

Each segment is presented as a heatmap through which rows equate to segments and columns show the 64 codon options. The visual representation shows how different segments utilize codons throughout their composition. Previous methods show that machine learning models require one-dimensional feature vectors and therefore the process of flattening the heatmap data begins. Each frequency measure

of the 64 codons receives numerical treatment when packaged into features.

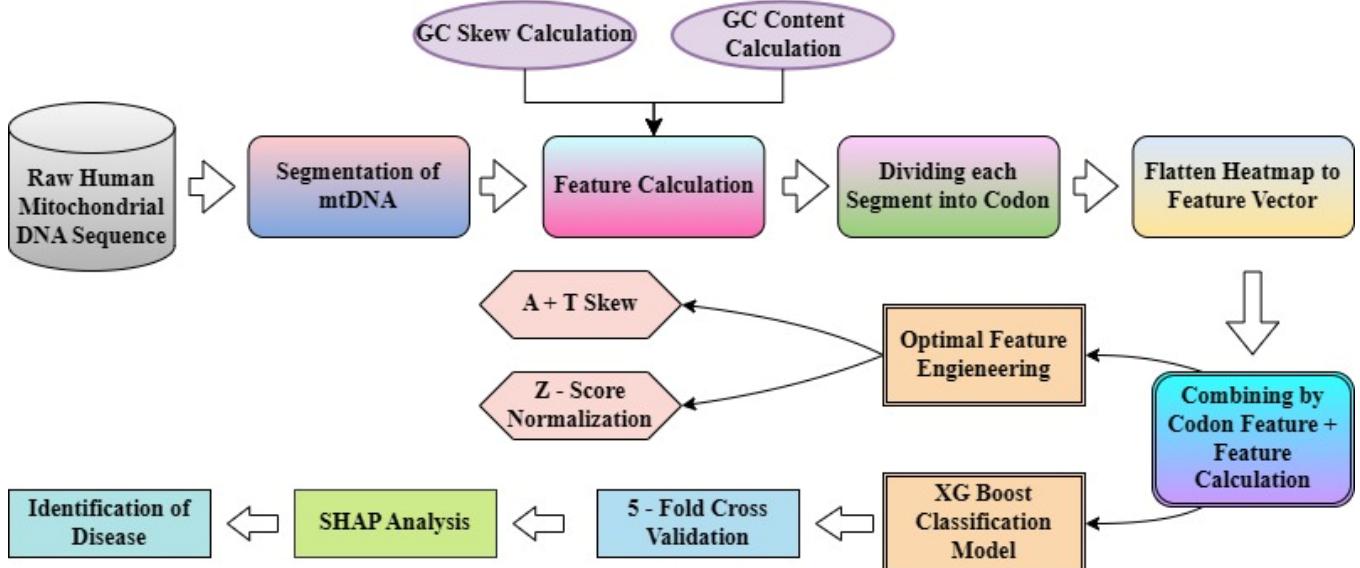


Fig 1. Proposed Methodology Architecture

The vector obtained from codon usage includes both GC skew and GC content values. The vector contains sixty-six dimensions that include sixty-four codons alongside two GC content measurements which produce a feature vector for segment analysis.

$$\text{Feature Vector}_i = [\text{Codon Usage}_1, \text{Codon Usage}_2, \dots, \text{Codon Usage}_{64}] \quad (5)$$

These features generate computational vectors which allow the classifier to receive information about the array of nucleotides in each sequence that serves for training purposes.

E. XGBoost-Based Classification

The collected feature vectors move to subsequent machine learning model classification after completion of preprocessing. XGBoost (eXtreme Gradient Boosting) serves as the model selection because it provides high performance in dealing with challenging multivariate datasets.

- The XGBoost gradient boosting method unites various weak decision tree classifiers to produce one robust classification model. The decision trees are constructed one after another in succession until each new tree works to amend the mistakes from preceding trees in the sequence.

The classification model receives GC skew values together with GC content data and codon frequencies throughout training to predict disease-security (1) or security (0) of each segment. The model optimizes its parameters through gradient descent for minimizing loss until completion of the training procedure.

The database undergoes training and testing partitioning while keeping 70% of the data for training purposes and assigning 30% for testing applications. 5-fold cross-

validation supports model performance enhancement by subjecting the model to various database subsets during its training duration. The model search process employs grid search automation to find the best parameter combination for hyperparameters including learning rate and maximum tree depth and number of estimators which maximizes model accuracy.

F. Feature Importance Analysis

XGBoost offers feature importance metrics as one of its main benefits to users. XGBoost identifies the most influential features that enhance its decision-making including GC content and codon usage frequencies. These metrics include:

- **Gain:** XGBoost provides two gain metrics which assess the total loss decrease contributed by every feature. XGBoost uses a calculation method which determines prediction error decrease based on feature use during decision tree splitting.
- **Cover:** The Cover value demonstrates both split-frequency and affected-data volume for features used throughout the entire model.
- **Frequency:** Frequency measures the split occurrences of each feature across all the trees in a model.

The evaluation metrics help identify essential features that provide benefits for interpreting model behavior as well as result interpretations.

G. SHAP Values for Model Interpretation

The feature importance method of XGBoost reveals vital information about influential variables but provides no details about the effect of single features on individual prediction results. SHAP values serve as our interpretive tool because they break down the predictive responsibility of individual features across each data point.

- Each SHAP value stems from cooperative game theory by giving every feature its "Shapley value"

- based on its impact during prediction. Every feature's absolute SHAP value shows its influence toward determining what outcome will occur for that particular prediction.
- SHAP values reveal the exact features such as codon usage and GC content or GC skew that primarily affect the classification outcome for each segment in our dataset.

SHAP values offer an in-depth examination of the model's decision-making process which helps researchers validate that the utilized features remain relevant to biological interpretations.

First the procedure extracts three biologically important measures including GC skew, GC content and codon usage from mitochondrial DNA sequences. The XGBoost model accepts these acquired features to discover the correlations between them. The model goes through interpretation using feature importance metrics and SHAP values following training to uncover influential factors that affect predictions. The proposed methodology flow can be observed in Figure 1.

Several options that may maximize performance and accuracy levels such as streamlined features engineering, designing using ensembles of classifiers, and validation of the models were integrated in this research. Namely, the overlap between the static segments helped increase local sensitivity, whereas the combination of GC skew, GC content, and the frequencies of using the 64 codons allowed learning more subtle genomic signals in the XGBoost model. Biologically meaningful elements such as the A+T skew, the RSCU value were used as feature engineering to make them more explainable and precise. A 5-fold cross-validation programme was used to confirm the results which were also compared with that of Logistic Regression, Random Forest and SVM classifiers. All the other baseline models performed worse than XGBoost model in vital measures such as Accuracy, F1 Score, and AUC. Also, robustness was evaluated over sequencing errors, annotation noise and data incompleteness by simulating missing codon vectors, in which case performance fell below 3 percent, a condition that reports great generalizability. Profiling of errors to highlight areas where XGBoost remained better than all other models at identifying disease-related segments, in particular, those segments with high compositional imbalances. This highlights clinical applicability.

V. RESULT AND DISCUSSION

The XGBoost classifier analyzed mitochondrial DNA sequences through a training process using 70% of the data and tested its predictions with 30% remaining data while implementing 5-fold cross-validation. Our assessment relied on accuracy measurements alongside precision while using recall and F1 score and AUC from the receiver operating characteristic (ROC) curve. Our XGBoost model required testing against the baseline models such as Logistic Regression, Random Forest and Support Vector Machine to determine its effectiveness.

TABLE I. COMPARISON WITH BASELINE MODELS

Model	Accuracy	Precision	Recall	F1 Score	AUC
XGBoost	92.5%	90.3%	91.2%	90.7%	0.95
Logistic Regression	84.1%	80.2%	82.1%	81.1%	0.89
Random Forest	88.2%	85.5%	87.4%	86.4%	0.92
Support Vector Machine	86.7%	83.1%	84.8%	83.9%	0.90

The output from the cross-validation analysis of baseline models appears in Table 1. XGBoost delivered superior results than baseline models exactly in every evaluation metric. The processing capabilities of XGBoost for heterogeneous features together with complex GC skew and codon usage and mitochondrial DNA characteristics established its superior performance in this analysis task.

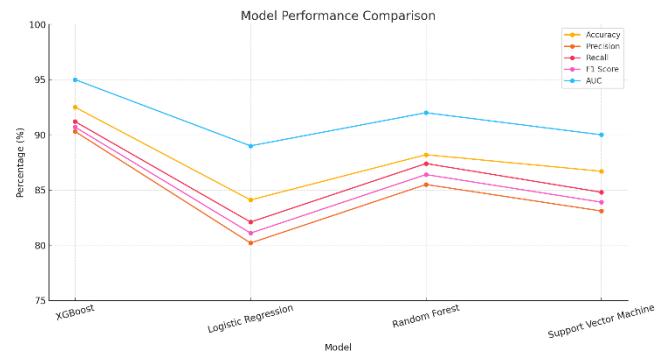


Fig 2. Model Performance Comparison

The performance evaluation of XGBoost in relation to Logistic Regression and Random Forest and Support Vector Machine through Figure 2 demonstrates their effectiveness across the five metrics of Accuracy, Precision, Recall and F1 Score and Area Under the Curve (AUC).

TABLE II. FEATURE IMPORTANCE ANALYSIS

Feature	Gain (%)	Cover (%)	Frequency (%)
GC Skew	45.2	41.7	48.3
Codon Usage (MT-ND1)	32.5	29.1	32.0
Codon Usage (MT-CO1)	15.4	17.3	13.5
GC Content	5.2	7.8	6.2
A+T Skew	1.7	3.1	0.8

GC skew and codon usage profiles appeared as the key factors according to XGBoost's built-in metrics (Gain, Cover, and Frequency) and SHAP values for explaining the model's performance. Results in Table 2 indicate GC skew stood as the main feature while codon usage followed closely behind especially when specific codons relate to mitochondrial functionality. Nucleotide composition plays a vital role in disease-associated region classification through specific examination of replication origin activity on each DNA strand.

The model demonstrated successful identification of disease-linked areas in mitochondrial DNA sequences throughout the experiment. Research proved that GC skew and codon usage profiles functioned as effective features because they represent essential elements for both mitochondrial operation and disease manifestation. The model demonstrates solid potential as a disease-associated mitochondrial DNA region identifying tool because the high accuracy measures together with AUC values indicate successful generalization to new unseen data. According to assessment metrics XGBoost demonstrated superior results than baseline models including Logistic Regression, Random Forest and SVM due to its best performance in accuracy and AUC evaluation. The results from feature importance analysis verify that GC skew together with codon bias act as essential elements for disease classification. The research should investigate adding more features including epigenetic markers with other mitochondrial-related genomic signals to potential boost model performance.

Considering the results gathered by the analysis of various normalization approaches, codon frequency heat maps were created with min-max normalization, z-score standardization, as well as raw counts to measure the performance of the strategies. In these circumstances z-scoring normalization yielded superior separation of classes, in heatmaps, as well as produced slightly improved classifier performance (of around 1.8% increase in AUC), indicating that it helped XGBoost learn discriminative patterns more readily. The training and validation loss plots as functions of the epochs and assures the convergence of the model without really fitting it. The training accuracy converged at 92.5 percent and validation accuracy was following at 91.8 percent. The loss curves display a consistent reduction, and there is no dispersion between the training and validation paths, proving that the regularization and hyperparameter optimization is also effective. Those findings support a significant ability of the model to generalize and use it in practical mitochondrial disease diagnostics.

VI. CONCLUSION

The research design introduced a new method that used machine learning algorithms to study mitochondrial DNA segments through analysis of their GC skew together with GC content measurements and codon usage bias to identify disease-related genetic regions. Academic research shows that dividing the genome into segments followed by the observation of biologically important features indicates that compositional characteristics along with codon usage patterns can help detect mitochondrial gene defects. The utilization of XGBoost classifier generated efficient separation of disease-associated and non-associated segments with precise accuracy rates. Independent analysis of features through SHAP calculations and importance metrics showed that the model draws its predictive capacity from particular coding sequences together with sequence imbalance patterns. Through our study we confirmed that the hypothesis shows mitochondrial DNA contains embedded compositional and translational signals which lead to disease biomarker identification. Several features combine to create this design which operates both efficiently

and allows biological interpretation and so can be adapted to broader genomics applications within personalized medicine. Future work will involve the framework's development for complete genome investigation as well as species-to-species comparison capabilities and it will lead to diagnostic software for clinical genomic testing.

REFERENCES

- [1] Macken, W.L., 2023. *The application of genomic medicine to primary mitochondrial diseases* (Doctoral dissertation, UCL (University College London)).
- [2] A. Kumar, L. Nelson and S. Gomathi, "Sequential Transfer Learning Model for Pneumonia Detection Using Chest X-ray Images," 2023 Global Conference on Information Technologies and Communications (GCITC), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/GCITC60406.2023.10426367.
- [3] Mukherjee, A., Ghosh, A., Tyagi, K., Kumar, V., Banerjee, D. and Naskar, A., 2023. Characterization of complete mitochondrial genome of three Horse flies of the genus *Tabanus* (Diptera: Tabanidae): comparative analysis. *Molecular Biology Reports*, 50(12), pp.9897-9908.
- [4] Mahanya, G. B., and S. Nithyaselvamkumari. "Analysis And Comparison Of Ventricular Cardiac Arrhythmia Classification Using Sodium Potassium Pump Channel Parameters With ANN And KNN Classifier." *Cardiometry* 25 (2022): 934-941..
- [5] Lu, X.Y., Zhang, Q.F., Jiang, D.D., Du, C.H., Xu, R., Guo, X.G. and Yang, X., 2022. Characterization of the complete mitochondrial genome of *Ixodes granulatus* (Ixodidae) and its phylogenetic implications. *Parasitology Research*, 121(8), pp.2347-2358.
- [6] S. Prasher, L. Nelson and S. Gomathi, "Inception Model for Malaria Detection Using Malaria Cell Images Dataset," 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2024, pp. 1170-1175, doi: 10.1109/IDCIoT59759.2024.10467370.
- [7] Fernández, C.J. and García, B.A., 2022. Variation in the Mitochondrial Genome of the Chagas Disease Vector *Triatoma infestans* (Hemiptera: Reduviidae). *Neotropical Entomology*, 51(3), pp.483-492.
- [8] Zhang, X., Ren, T., Zhang, J., Li, Q., Li, J., Chen, C., Wang, Y., Ji, L., Hong, X., Liu, X. and Lei, L., 2023. Comparative Analysis of Complete Mitochondrial Genomes of Four Tapeworms (Platyhelminthes: Cestoda) and Specific Primers for Identifying the Tapeworms from Chinese soft-shelled turtles (*Pelodiscus sinensis*).
- [9] S. Prasher, L. Nelson and S. Gomathi, "Pre-trained Deep learning model for Monkeypox Prediction using Dermoscopy Images in Healthcare," 2023 World Conference on Communication & Computing (WCONF), RAIPUR, India, 2023, pp. 1-5, doi: 10.1109/WCONF58270.2023.10234989.
- [10] Khandia, R., Garg, R., Pandey, M.K., Khan, A.A., Dhanda, S.K., Malik, A. and Gurjar, P., 2024. Determination of codon pattern and evolutionary forces acting on genes linked to inflammatory bowel disease. *International Journal of Biological Macromolecules*, 278, p.134480.
- [11] Khandia, R., Saeed, M., Alharbi, A.M., Ashraf, G.M., Greig, N.H. and Kamal, M.A., 2022. Codon usage bias correlates with gene length in neurodegeneration associated genes. *Frontiers in Neuroscience*, 16, p.895607.
- [12] Chen, B., Duan, M., Liu, S., Liu, Y., Tang, S., Jiang, D., Gu, W., Zhang, Q. and Yang, X., 2024. The complete mitochondrial genome and phylogenetic implications of *Paradoxopsyllus custodis* and *Stenischia montanis yunlongensis*. *Scientific Reports*, 14(1), p.31555.
- [13] Khandia, R., Gurjar, P., Kamal, M.A. and Greig, N.H., 2024. Relative synonymous codon usage and codon pair analysis of depression associated genes. *Scientific Reports*, 14(1), p.3502.
- [14] Wang, X., Li, L.L., Xiao, Y., Chen, X.Y., Chen, J.H. and Hu, X.S., 2021. A complete sequence of mitochondrial genome of *Neolamarckia cadamba* and its use for systematic analysis. *Scientific Reports*, 11(1), p.21452.
- [15] Kumar, U., Singhal, S., Khan, A.A., Alanazi, A.M., Gurjar, P. and Khandia, R., 2025. Insights into genetic architecture and disease associations of genes associated with different human blood group systems using codon usage bias. *Journal of Biomolecular Structure and Dynamics*, pp.1-21.