A Survey on Genomic Dataset for Predicting the DNA Abnormalities Using Ml

Siripuri Divya[1], Y. Bhavani[2], Thota Mahesh Kumar[3]
M.Tech, Data Science, Kakatiya Institute of Technology & Science Warangal, India[1]
Associate Professor, Dept. of IT, Kakatiya Institute of Technology & Science, Warangal, India[2]
Assistant Professor, Dept. of IT, Kakatiya Institute of Technology & Science, Warangal, India[3]

***Abstract*—Genomic data is used in bioinformatics for collecting, storing and processing the genomes of living things. In order to process the genetic information, machine learning algorithms plays a vital role in building a computational model by using the statistical theory. This paper helps the researchers, who are doing research with the DNA dataset by applying the machine learning logics. Feature scaling machine learning techniques helps in predicting the sequence of genome for extrachromosomal amplification and predicting the tumor intensity in the human gene. Identification of unconventional chromosome in the DNA sequence minimizes the structural risk. In this paper, researchers can get clear insight on classification, sequence prediction, fuzzy relationship and SNP on genome dataset. The performance of various existing models is measured using the performance metrics and the accuracy.**

***Keywords*—Genomic data; deoxyribonucleic acid (DNA); machine learning algorithms; single nucleotide polymorphism (SNPs)**

I. Introduction

Deoxyribonucleic acid (DNA) is the combination of different nitrogenous bases, phosphate molecule and sugar molecules which are combined to form nucleic acid. These nucleic acid molecules consist of genetic information that are used for transmitting organic material from parent to child. All the combinations in the nucleic acid of a cell are arranged in a sequence to form a DNA structure. DNA are mainly responsible for storing genetic information and also for producing the proteins in human body through transcription and translation process. DNA are in double helix structure, present in eukaryotic and prokaryotic cells.

Deoxyribonucleic acid (DNA) is located in each and every cell nucleus of human body containing the genetic information. The cell nucleobases of DNA consist of four nitrogen nucleotides namely adenine represented as A, cytosine represented as C, guanine represented as G and thymine represented as T formed using the nitrogen bond as represented in the Fig. 1. DNA of a cell nucleus consist of chromosomes. There are 3-billion pair of chromosome sets termed as genome. Genome consist of 46 chromosomes in the DNA sequence grouped to 23 pairs.

DNA can be divided into four different categories based on the structure as A-form, B-form, C-form and Z- form DNA. The base pairs which are not in perpendicular to the helix axis are defined as the A-form DNA, these protect the human body in the extreme desiccation effect of bacteria. The base pairs which are perpendicular to the helix axis form are defined as the B-form DNA, these are responsible for gene expression, mutation under normal conditions. The zig-zag form of base pairs to the helix axis are defined as Z-form DNA, these are responsible for gene regulation. The base pairs which are in the form of non-integral helix structure are defined as the C-form DNA, these are responsible for cloning the eukaryotic genes.

DNA consisting of 3 billion base pairs carrying information from one gene to another gene are created and replicated based on the human genetics and the life-style. The DNA arranges the base pairs in a chain sequence to perform transcription and translation process to produce protein in mRNA to develop the human body. These proteins are used as energy to do large amount of work by our body. The DNA structure is divided into two halves or two single strands. The strands in DNA are thin molecules wrapped to form a helix structure, these are composed of blocks or nucleotides. Human genome is around 98% similar to chimpanzees and 75% similar to mouse.

Fig. 1. Deoxyribonucleic Acid [23].

A. DNA-Sequence-Classification

DNA can be classified based on various standards like structure, number of base pairs, location, coiling patterns, nucleotide sequences, number of strands, coding and non-coding. DNA sequence-classification is performed based on the nucleotide combinations in the sequence. The four nucleotides A, T, C and G forms a series in the nucleic acid of the cell. The process of identifying the nucleotide sequence in the DNA are used to determine the order of nucleic acid sequence. DNA sequences are classified to get the information related to the evolution of various species, living organisms and their transformations, medical-diagnoses, forensic

investigations and organism identification. The nucleotide canonical DNA structure using the computer terminology can be classified for various predictions like disease risk, next generation sequence, cancer research, birth defect screening, food importing/exporting control, paternity testing, drug target and gene therapy.

The process in which the information from the DNA is carried out to the RNA molecules are called as transcription. Transcription process is carried out in three different stages in the gene expression, enzymes that perform transcription in copying the DNA strand from the DNA [24] sequence in eukaryotes cell is called RNA polymerases. The single stranded DNA correlation with the complementary strand RNA is performed using the RNA polymerases by adding the new-nucleotides. The first step in the gene transcription is initiation of promoters for binding the RNA polymerase with the DNA sequence molecules in each gene. The second step is elongation process where the RNA molecule builds the complementary nucleotides chain in the template strand. Termination is the final step in the transcription where the sequence mechanism is formed as hairpin RNA molecule.

The process in which the information from the termination of RNA molecule is translated to the ribosome to produce proteins is called translation process. In the translation process the genetic code from the RNA molecule is converted to the amino acid 20-letter code to produce the protein blocks.

The translation in the ribosome is performed in three different stages. The first step of gene translation is initiation, where the small-ribosomal-subunit binds the information from the transcription and codons initializes the methionine code & AUG to transfer the information. The second step of translation is elongation where the codon continues to increase the chain by adding the corresponding amino-acid using the peptide bond. The final step in the translation process is the termination where the proteins are produced by completely binding the codons from the RNA molecule.

Fig. 2 displays the transcription control flow and the translation process, in converting the DNA information to produce the proteins. Transcription synthesis of single stranded RNA from a double stranded DNA template is used to produce messenger RNA. Translation is the first stage of protein biosynthesis from RNA in the gene expression.

B. DNA Methylation

DNA Methylation is the process in which the methylation activity in the DNA segment is changed without changing the original gene sequence. Methylation is a process where the methyl groups are attached to the DNA molecule to repress-gene. DNA methylation occurs during the epigenetic event. The covalent modification of DNA methylation results in three types of methylated bases called C5-methylcytosine(5mC), N4-methylcytosine(4mC), N6-methyladenine(6mA). The DNA methylation is important for transcriptional-gene-silencing, genomic imprinting, maintaining the genome stability, embryonic development and X-chromosome inactivation's.

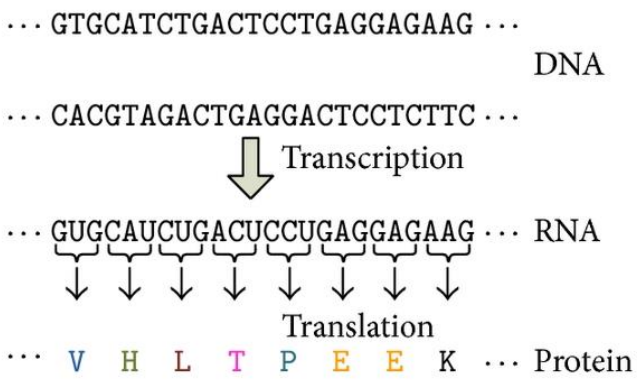

Fig. 2. DNA Transcription and Translation.

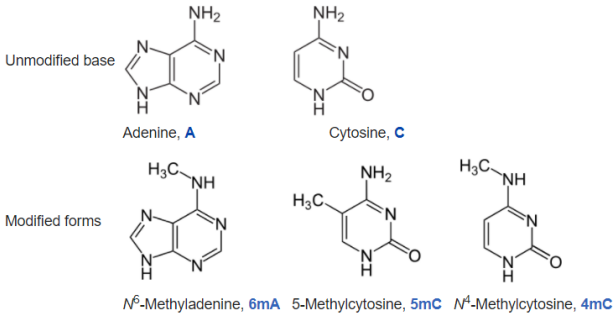

Fig. 3. DNA Methylation.

Fig. 3 is the chemical modification of the cytosine, where by addition of a methyl group to the number 5 carbon of the cytosine is converted to 5-methylcytosine which is followed by the guanine dinucleotide CpGs process.

C. DNA Damage

DNA damage is the process of alteration in the DNA structure resulting in the chemical abnormalities. DNA damage are mainly caused due to the change in the environmental factors and the metabolic process inside the cell. The main source of the DNA damage is endogenous damage with in the cell and exogenous damage caused by the external agents like X-rays, UV-rays. The DNA damage can be classified into three different types based on the alteration in the genetic material as single-base-alteration, two-base-alteration and chain-breaks-and-cross-linkages. The abnormality in single base of the DNA is caused by depurination, alkylation, deamination and base-analog formation. The two base alteration is caused by UV induced dimer formation in the thymine and bifunctional-alkylating agent. The chain-breaks-and-cross-linkages are caused by ionizing radiation, oxidative-free-radical-formation, radioactive disintegration, cross linking-between bases in same or opposite strand, cross linking between DNA and protein molecules. The Fig. 4 represents the DNA damage caused during the cellular alteration. Agents damaging the DNA [22] are radiations caused due to highly reactive oxygen radicals, ultraviolet rays and ionizing radiations, chemicals in the environment like aromatic hydrocarbons and aflatoxins.

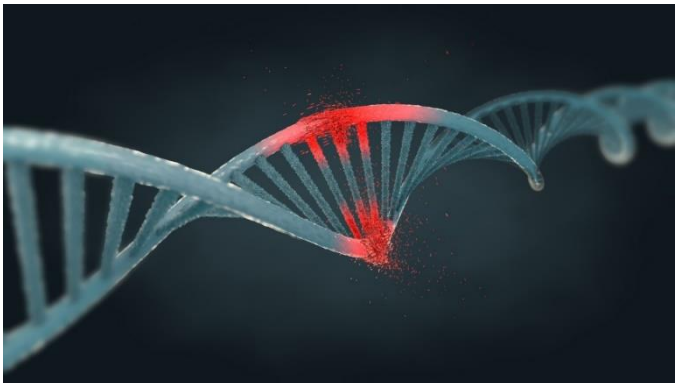

Fig. 4. DNA Damage [26].

DNA repair can be performed by the cell in two ways, direct-damage-reversal and excision of DNA damage. In direct-damage-reversal initially the single polypeptide with the enzymatic properties binds the chain and restores using DNA photolyases and alkyl transferases. In the excision of DNA damage, it solves the excised free bases generated by altering the bases to deoxyribose-phosphate by initializing the DNA glycosylases called base-excision-repair. Nucleotide-excision-repair mechanism is used to replace the DNA damage in 30 bases. Mismatch repair mechanism allows the enzymes to identify the strand and replace them with normal cellular enzymes corresponding to the base-pair-rules, strand break repairs, single-strand breaks and double-strand break damages. The diseases caused due to the defect in the DNA repair system are ataxia telangiectasia, bloom syndrome, Cockayne's syndrome, progeria syndrome, rothmund-thomson syndrome, trichothiodystrophy, Werner syndrome, xeroderma pigmentosum and hereditary non polyposis colon cancer.

Fig. 5 represents the Nucleotide-excision-repair mechanism is used to replace the DNA damage in 30 bases.

D. Mitochondrial DNA

Mitochondrial DNA is responsible for the cellular-metabolism, oxidative-stress-control and apoptosis. Mitochondria is also known as the power-houses of DNA cell, it is inherited from the mother's ovum. It consists of 13 coding genes and the 24 non-coding gene of length 16,569 bp in the human body. Mitochondria uses the oxygen and sugars to create energy of the main cell. The mitochondria mainly designed with 22tRNA and rRNA coding genes to control replication and transcription process in the cell. It is in circular structure with several copies of single mtDNA-molecules present freely in the nuclear envelope.

Fig.6 Mitochondrial organelles structure are found in the cell cytoplasm as represented in Fig. 6 which consist of some of components like 2 membranes called inner membrane and outer membrane protecting the cellular matrix.

E. DNA Mutation

The heritable change in the arrangements of genetic material chromosomes position is termed as mutation or DNA mutation. Mutations are occurred in gametes and causes permanent change in the genetic sequence of nucleotide forming new amino acid. The gene mutation rate is 1 or 2 new mutations in 1000000 genes during the DNA copy. Mutations are unpredictable and can be of many forms like gene mutation or point mutation and chromosome-mutation. These mutations are caused due to change in the complete chromosomes structure, chromosomes count and single pair of chromosome's structure. Some of the forms of mutations are due to the addition of extra nucleotide that causes gene mutation and addition of extra chromosome that causes chromosome mutation. Deletion of nucleotide chain from the gene sequence causes gene mutation and chromosomes are lost in the gene sequence causing chromosome mutation. Duplication of nucleotide chain is repeated in gene mutation and chromosomes are repeated in chromosome mutation. In inversion nucleotide sequence are detached from the gene sequence causing gene mutation and deleted chromosomes rejoining the chain in the inverse position causing chromosomes mutation.

Fig. 7 represents different mutations caused in the DNA. The first sequence in the Fig. 7 represents the normal gene sequence consisting of cytosine, thymine, adenine and guanine. The insertion of new guanine in the sequence at the second position results in the mutation. The deletion of adenine in the 3rd position of the sequence, duplication of cytosine and thymine, inversion of the 2nd and 3rd position of the sequence causes DNA mutation.

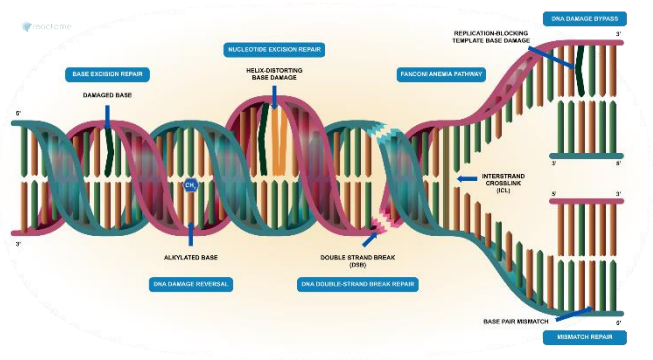

Fig. 5. DNA Repair [25].

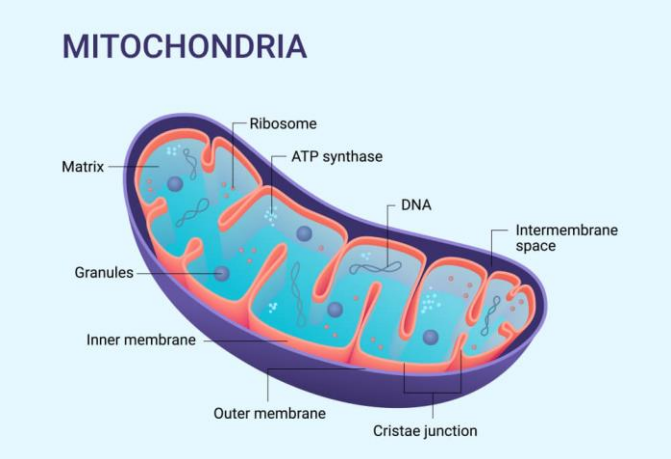

Fig. 6. Mitochondria.

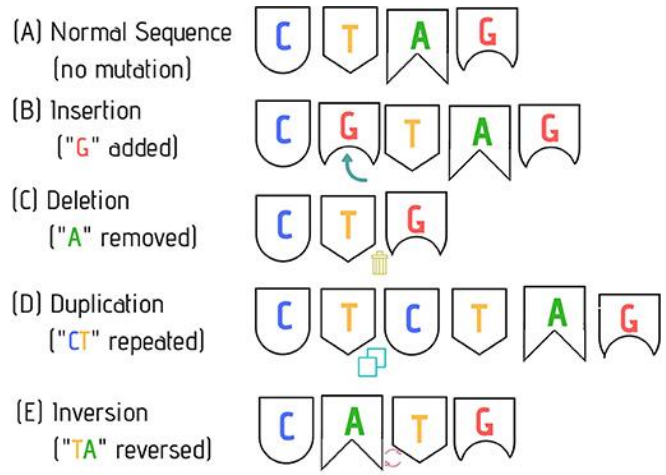

Fig. 7. DNA Mutation.

II. Background and Motivation

Deoxyribonucleic-acid composed of 46 chromosome's carries genetic information from parents to children in homo sapiens. Chromosomes are made up of different groups of purine-pyrimidine bases, phosphate and sugar attached to the one carbon of deoxyribose's forming as adenine represented as *A*, cytosine represented as *C*, guanine represented as *G* and thymine represented as the *T* formed using the nitrogen bond. Human genome is used for scientific research for understanding the physical gene sequence and base pairs functioning. DNA research for analyzing and classification of human genome to predict the accuracy of disease effecting a person through machine learning approaches became enormous in the present-day. Research in DNA is unearthing the evolution of human in the nature by mapping the gene based on the physical features. DNA classification of chromosomes can be used for predicting numerous inheritance mutations and DNA damages. Human-Genome-Project which was started in 1990 and continued till 2003 worked on the base's sequences in the gene sections of 3 billion long with an average size of 3000 bases. Human-genome-project work done by scientist all over the world improved the medical field, microbial genome research, forensics, disease risk assessment and human evolution. DNA abnormalities can lay a step forward for predicting the multifactorial inheritance risk in effecting children from parents.

III. Related Work

The classification of eukaryotic genome linear DNA chromosomes on extrachromosomal DNA by Zhenyu liao et al., in 2020 [1] provided a circular structure of the extra chromosomes which were found outside of the eukaryotic-genome. The authors had classified the unconventional chromosomes for identifying the cancer tumor miscellaneous behavior in the gene. The extrachromosomal were classified based of the spectrum of microscopy technology for detecting the progression of tumor. According to the authors research the extrachromosomal DNA is mainly characterized into four types based on the size of the chromosomes, frequency in the tumor cells and functionality. The progression in the tumor were relatedly close to two gene amplification in the chromosome segmentation which increases the intensity in the tumor cells. The authors discussed about the drugs that resist to the tumor cells and formulated the extrachromosomal cycle and its proliferation in the genome. The extrachromosomal DNA formulation gives the complete information regarding the circular chromosome amplification and its translocation of tumor cells. The super resolution of the next generation genome sequencing for different elements of the extrachromosomal DNA tumor identification and amplification were resolved using gene editing tools.

Receiver-operating-Characteristic-curve was achieved by Leif E.Peterson et al.,[2] for cancer microarray DNA using different Machine-learning classifiers, feature-scaling and fuzzification for the 9 cancer microarray datasets. The sample size is initially considered for obtaining the AUC values for the fuzzy set and the crisp set based on the statistics to recommend the factor which is influencing the AUC percentage value. The inferential hypothesis test is used for predicting the effect on the AUC value. The feature-scaling of the dataset is performed using the t-test suboptimal ranking for N inputs. The fuzzy logics were used to get the real gene expressed in the cancer microarray dataset. Machine learning logics were used to find the regression and classification of the cancer microarray datasets using certain formulas for supervised classification. The authors had compared the results of both the fuzzy and crip accuracy of the 9 datasets in the graphical representation and the cancer microarray are categorized based on the classification of the data. The AUC fitting for the cancer datasets for the least and the highest correlations are obtained based on the feature-scaling and fuzzification performance.

Stephen winters-hilt et al.,[3] had developed a new computational method for Single-Molecular DNA Classification to enhance the accuracy of hairpin DNA using single species data in silico. This computation method was named as Watson-crick-basepairs. The authors in this model, performed the SVM multiclass architecture to analyze the state of DNA molecule and also its transitions in the molecular structure with respect to the kernels. The Hidden-Markov-model parameters are also used for denoising and feature-vectors in the molecule. The model performance is measured for a single DNA molecule using the performance metrics in the biophysical analysis. Nanopores are generally used to measure the DNA molecules for each and every basepairs, the feature extractions are performed from multiclass scalability trends to analyze the sequential data of basepairs. SVM provides the optimized hyperplane which separates the hyperplane into clusters for mapping of feature-vector to discriminate the structural risk in DNA hairpin in the experimental procedure. This model achieved a highest accuracy of 99.6% in less than six seconds.

A new method for handling classification problem in the DNA coding was proposed by Ting-Cheng et al., in 2015 [4] as variable-coded-hierarchical-fuzzy-classification-model (VCHFM). The supervised learning method is used as an interface for fuzzy system and the DNA coding. This model works on four main principles. The first main principle of VCHFM is automatic fuzzy rules generation for numeric data and feature-extraction. The second principle works on the DNA computation functions. The third principle works on the

optimization of the chaotic particle that regulates the weight grade of the inference node in DNA. The final principle works on classification functions and the multi-objective-fitness optimization function. This model is highly capable of reducing the overlapping problem and dimensionality reduction problem that affects the classification. VCHFM obtained a benchmark result in best classification rate with a smaller number of fuzzy rules.

Umit Atila et al., in 2020 [5] had classified DNA damage problem using convolutional neural network using the comet images of the grayscale DNA. The authors had worked on the quantification of the images and the identification of the damage comet object in the DNA. They divided the entire DNA into four categories namely healthy, poorly defective, defective and very defective based on this the images were classified in the neural network. Comet-assay-experiment was conducted to obtain the images of 170*170-pixel resolution images for four categories labeled as G0, G1, G2 & G3. Authors had achieved a highest accuracy of 96.1% in predicting the damage DNA using convolutional neural network.

R. Touati et al., in 2021[6] provide a detail description on converting the DNA sequences into the chaos-game representation using the FCGR images. Authors had initially considered the helitron-family FCGR images for feature extraction in the automated system to develop the DNA sequences of helitron-family. The authors also applied the machine learning methods for classifying the images of DNA using SVM, PTDNN and Random-forest algorithms. SVM is used to minimize the structural risk in the DNA by dividing it into different clusters in the hyperplane. Pre-trained deep neural network is used for classifying the DNA images using the softmax activation function on the 2D images to classify the images. Random forest techniques like bagging are used for detecting the different variation of DNA images. The accuracy of classification of the DNA with all the three machine learning approaches were analyzed.

L. Liu et al., [7] provided the detailed description of the cancer plasma cell detection from the methylation sequence and its classification. Authors had developed a comprehensive-methylation-sequence by targeting a single plasma and identified the presence of cancer. The molecular testing method is used for classification of the cell-free DNA in the cancer gene to detect the defected plasma.

Sergio Bittanti Simone Garatti Diego Liberati[8] had provided a detailed description of the degeneration effect in the various types of cancer using unsupervised clustering. Author had classified leukemia and paradigmatic using data mining techniques. To analyze the data the authors had used the microarray technology for training the gene expressed data and for performing unsupervised clustering for diagnostic of DNA. Using this approach authors are capable to classify the data without any pathological information.

SNP-Single nucleotide polymorphism impact on DNA is classified by the Jard H. de Vries et al., 2021[9] which is used to solve the investigation of a murder case from the genomic data. The SNPs are generally used for obtaining the DNA quality & quantity, in the crime case by applying the global-screening-array to the sample. The impact of SNPs is used to classify the kinship, based on the positive and negative values of kinship-classification the murder case is solved.

Jun Hu et al., in 2020 [10] provided information related to the protein sequence analysis from DNA-binding using the computational methods. Authors had worked on the target DNA-binding protein by applying four feature extraction operations. The base features of the DNA are extracted using the Amino-Acid-Composition, Pseudo-position-specific-scoring-matrix, pseudo-predicted-relative-solvent-accessibility and pseudo-predicted-probabilities-DNA-binding-sites. They had combined both the base features and their weights to determine the original-super-feature to perform the machine-learning algorithms. Authors had used statistical predictors to improve the accuracy of the DNA-Binding-protein, dataset analysis, feature extraction, multi-view-features and feature selection to identify the DNA-sequence. The results of applied feature selections are measured using the performance metric like ROC, accuracy and CBR ranking.

The cancer classification based on the DNA-mutation patterns were performed by the Lei Wu et al., in 2020 [11]. The amplifications in the tumor cells are identified in the DNA-sequence using the Surface-Enhanced-Raman-Spectroscopy (SERS). The authors had created a free amplification SERS sensor to integrate it into microfluidic-chip with in the DNA-nucleotide mixture for demonstrating the melanoma-cell lines and colorectal cancer. The SERS classification of cancer types using the profiling mutation in the DNA patterns had achieved a benchmark accuracy of above 90%.

The classification of DNA-microarray using advanced algorithms was discussed by Beatriz A. Garro et al., in 2016 [12] for synthesis of mRNA molecules. The authors had diagnosed various diseases to identify the tumor and detect its amplification in the cells. The new algorithms using artificial-neural-network to solve the classification problem in the DNA groups associated for a particular disease in gene expression. The computational models used by the authors to classify are artificial-neural-network, multilayer-perceptron, radial-basis-functions and support-vector-machine. In the feature selection process, the authors had evaluated the accuracy of the model by using four different datasets for a particular disease using ABC algorithm.

Firoz khan et al., in 2020 [13] provided the information related to the digital DNA-sequencing to detect the ransomware using various machine-learning approaches. Software used to implement digital DNA-sequencing is ransomware, it is one of the classes of malicious software which is used to predict various attacks over the internet. Author had developed appropriate ransomware attack flow with machine-learning algorithms called DNAact-Ran (engine) methodology to predict the DNA-sequence. The DNAact-Ran engine is evaluated using the machine-learning performance metrics for analyzing the accuracy and engine effectiveness on real-time dataset.

The DNA-Binding model EL_LSTM was proposed by Jiyun Zhou et al., in 2020 [14] for residue relationship prediction. In this novel approach the authors had mainly

concentrated on the two concepts initially finding the pairwise relationship with long-short-term-memory bigram model and secondly solving the data-imbalance problem in DNA-binding. The authors considered four datasets namely PDNA-224, DBP-123, HOLO-83 and TS-61. The residue data instances are calculated using the sequence length and sub sequence length of the chain with protein sequence. Using the LSTM method, the experiment was conducted and the performance of each and every residue data instances are evaluated for neural-network, random forest, support vector machine and LSTM and accuracies of all algorithms are analyzed for all the four datasets.

Wunsch algorithm was proposed by Amr Ezz El-Din Rashed et al., in 2021 [15]. Biological-sequence-alignment-algorithm problems were addressed by authors using the Wunsch algorithm. In this model the DNA-sequence is considered as the input to the parallel workflow model, all the input sequences are performed with the machine learning models to decode the sequences label and to obtain the results. The traditional sequential model discussed in this paper are based on the sequential workflow consisting of input sequence, initialization of matrix, matrix score, traceback of matrix score and results generation. The Wunsch algorithm is capable of computing and converting the DNA-sequence from alphabet to the decimal or binary representation. The proposed model achieved a benchmark accuracy of 99.70% to prevent overfitting problem in the DNA-sequence.

Transcription-factor-binding was implemented using convolutional-neural-network by Qinhu Zhang et al., in 2021[16] for understanding the various DNA cellular-functions and binding mechanisms.

Erfan Aref-Eshghi et al., in 2018 [17] provided a detail description on monotonous debate between histone adjustment DNA methylation proposes that the aliment might be anticipated to show DNA-methylation impressions that ponder those primitive error related with chromatin myopathies. Here we study 14 mendelian states that show from direct disordering or indirect disordering of the proteins. To recognize Genomic regions, bear methylation. Switch, a jolt coursing approach is used by the jolt courser package. Ten-fold cross checking of this representation showed an accuracy of 99.6%. If exactly detected the class of the 141 pompous subject that are used in training with other samples obtainable from alike conditions & other diseases are being reviewed for discovery of epi-signatures.

Antonino Fiannaca et al., in 2015 [18] suggested adjustment free procedure for DNA barcode categorization that is established on both a phantom representation and a neural-gas-network (NGN), for unsupervised-clustering. Best results can be acquired using adjustment -free approaches established on phantom sequence representation. The main aim of the suggested process that is pondered an addition of our earlier work is the categorization of an unrevealed DNA sequence utilizing the frequency of a little set of k-mers. Here other classifiers reached almost (99%-100%) accuracy, but CT method reached 97%. The efficiency decays to approximately 95% at the family, species, and genus level with suggested method 98-96% accuracy is achieved when for analysing full-length sequences the score decays to 99-97% compared with suggested method.

Sara Alghunam et.,al. in 2019 [19] report on machine-learning that can be used for categorization, handling each dataset individually and connecting them. Micro-array is one of the fastest growing technologies in genetic research. When SVM and logistics were used they showed 50% and 45% respectively after feature selection process is applied, they showed 75% and 63% respectively. Spark & Weka libraries - when analysed spark libraries showed high accuracy and SVM (in Weka) has exceeded the other classifiers. Comparisons showed that GE data exceed DM, and SVM has given 99.68% efficiency.

Wenbin Liu et al., in 2020 [20] proposed a new miscellaneous learning review on ASM-SNP data (bi-polar disorder & schizophrenia). The recognized genetic differences in ASM-SNP data are key to disclose the underneath process of mental problems. Here the authors labelled the immediate confront via the latest miscellaneous learning and machine learning. New SNP (feature-selection) and continuous pathway-selection was examined and various miscellaneous learning meths included kernel-PCA, LLE...etc. The comparison from constitution clustering and imaging suggested that the misclassification between schizophrenia & Bi-polar disorder could be unavoidable for physicians. They achieved highest performance when using t-SNE and needed only 20% SNPS (Top-ranked) to achieve the best diagnosis.

Giulio Pavesi et al., [21] investigating whether practical details about genes can be predicted by using details obtained from their sequences combined with gene expression data. The SVM plays a crucial role in responding to the main request emerged from the work. To evaluate the categorization execution of SVMs authors worked on various holdout mechanism. They cautiously reviewed the election bias issue. For every training set obtained from proponent a subgroup of motifs outcome calculated by weeder. In specific, regarding one cluster, the genes percentage in test sets ranged from 68% to 70%. Experimentation as well case studies such as these may assist to shack more light on the present problem which remains amongst the most pertinent and calculated in bioinformatics and molecular-biology.

DNA functionality method that used the p53 malfunctioning in identifying the diseases [27] was proposed by Mikael S.Lindstrom et al., in 2022. The recognition of p53 worked was focused from many years with the enhancement in the biomedicine. Malfunctions associated with each disease can be identified using the clear insights of p53 malignancies. Authors used p53 for treating the cancer patients based of the DNA replicas. The fundamental procedure used for treating the cancer patients are p53-centeres multifaceted pathways and ribosome-biogenesis (RiBi). Author using the DNA replica and RiBi methods proposed a new approach which firstly deals with p53 canonical interaction and their regulation in post-translation of the target. Secondly, the response in DNA speed in performing replication with the p53 cellular genomic association with targeted cancer cells are given brief description. Emerging of p53 in replication stress (RS) are highlighted from the emerge of p53 to the key role in DNA

link replication. In addition, the tantalizing crosstalk in identifying the mediated monitoring between DNA replication and cell nucleolar RiBi are analyzed. Cancer diseases are outlined using the IRBC and the RiBi pathway tumorigenesis identification using the p53 malfunctioning in human. The p53 role in identifying the DNA replication and ribosome-biogenesis in cell homeostasis provided a clear vulnerability in identifying the cancer elucidation.

Prognostic investigation on DNA [28] methylation to identify the subtypes of tumors are proposed by Christopher et al., in 2022. Aberrant analyses of human DNA methyl patterns helped in identifying the subtypes of cancer diseases based on the response and outcomes. Authors used osteosarcoma malignancy procedure to perform chemotherapy using DNA methylation analyses. Authors worked on predicting the patient tumor with the help of the genomic methylation to identify the situation in the early stages. The patient response behavior to surgical reactions is also predicted using the hypomethylation procedure which derived high perfect outcomes. Downstream analysis for identifying the methylation patterns were performed in an experimental analysis using three datasets to derive site-specific methyl patterns. The experimental analysis was associated with the clinical human genomic outcomes.

Impact of mitochondrial DNA (mtDNA) in human brain postmortem detailed description was provided by Alba Valiente-palleja et al., in 2022. The authors [29] investigation pm mitochondrial DNA reveals the facts on the heterogeneous disorder genes that synthesize the phosphorylation oxidative systems. Neuropsychiatric symptoms are used for understanding the disorders of the human brain functioning. authors provided an empirical study on human brain tissues to alert the ageing process investigation in unequivocally diseases. The experimental analysis of this procedure on testing with various samples resulted a benchmark outcome in identifying the disorder using the mtDNA for finding the contradictory cells in human brain

IV. ML Approaches

In order to process the genetic information, machine learning algorithms plays a vital role in building a computational model by using the statistical theory. Table 1 describes the accuracy of various machine learning algorithms applied on different genome datasets. It helps the researchers in analyzing the DNA and its replications to various diseases and ML helps to identify the abnormalities in using the experimental procedure and optimization techniques to derive the accurate outcomes.

TABLE I. Accuracy of Various Machine Learning Algorithms Applied on Different Genome Datasets

DATASET	ALGORITHM	ACCURACY
Helitrons database	Pre-Trained Deep Neural Network (PTDNN) classifier	72.6%
Helitrons database	Support vector machine (SVM)	68.7%
Helitrons database	Random forest (RF)	91%
TF Binding Datasets	Deeper CNN	
UCI Pima Indians Diabetes	Fuzzy rule-based classification	73.70%
Glass	Fuzzy rule-based classification	60.04%
Wisconsin Breast Cancer,	Fuzzy rule-based classification	91.21%
Wine	fuzzy if-then rules.	99%
Iris datasets	Fuzzy rule-based classification	96.67%
PDB database	SVM-REF+CBR; without feature extraction	78.85
PDB database	SVM-REF+CBR; with feature extraction	79.71
Multiple datasets	Needleman–Wunsch (NW) algorithm	85.9
WDBC directory	Sequential minimal optimisation (SMO),k-nearest neighbour(KNN) ,and decision tree(BF-tree)	96.19%
WDBC directory	Hybrid of k-means and SVM	97.38%
The gene expression omnibus(GEO)	novel graph-based semi supervised learning algorithm	24.9%
National center of biotechnology information (NCBIGEO)	SVM and logistic regression	>75%
Orange laboratories	SVM and logistic regression on sparks	75%
Epsilon dataset and GECCO dataset	SVM logistic regression, and Naive Bayes on spark	>75%
NCBI GEO	SVM	<70%
PDNA-224	EL_LSTM	82.59
DBP-123	EL_LSTM	81.44
PDNA-224	LSTM	78.36
DBP-123	LSTM	80.51
PDNA-224	NN	72.34
PDNA-224	Rf	75.27
PDNA-224	SVM	74.98
PDNA-224	LSTM	78.36
DBP-123	NN	76.36
DBP-123	Rf	77.29
DBP-123	SVM	78.34
DBP-123	LSTM	80.51
Ransomware	Multi-Objective Grey Wolf Optimization (MOGWO)	78.5%
Ransomware	Binary Cuckoo Search (BCS) algorithms.	83.2%
ALL-AML	MLP	1.0000
ALL-AML	SVM	1.0000
ALL-AML	SVM	1.0000
ALL-AML	KNN	0.9736
ALL-AML	SMV	0.9583
ALL-AML	KNN	0.9412
BREAST	SVM	1.0000
BREAST	SVM	0.9470

BREAST	J48	0.9381
BREAST	SMV	0.8421
PROSTATE	MLP	1.000
PROSTATE	SVM	0.9804
PROSTATE	SMV	0.9706
PROSTATE	LDA	0.9550
PROSTATE	LDA	0.9118
BOLD database	SVM	64.8%
Comet assay database	CNN	96.1%.

V. Current Challenges

The researchers, who are doing research with the DNA dataset by applying the machine learning logics can work on this current challenges.

- Improvements in the medical field.
- SNP for forensic investigation.
- DNA samples for identifying the human genetics.
- Inheritance disease identification.
- DNA damage in the gene sequence.
- DNA sequencing accuracy enhancement.
- DNA degradation.
- DNA analysis for body fragmentation.
- DNA mutation identification.
- DNA affect in health development.
- DNA profiling.
- DNA analysis with the high-end technologies.

VI. Conclusion

This paper helps the researcher those who are doing research with the DNA dataset to choose the appropriate machine learning logics depending on the model accuracy. They are given scope for predicting the sequence in the genome for extrachromosomal amplification identification using the feature scaling. By applying machine learning techniques in the unconventional chromosome can minimize the structural risk. DNA mutations can also be predicted using the genome data analysis using the statistical theory. Experimentation as well case studies shown in this paper may assist to shack more light on the present challenges.

References

[1] Liao, Zhenyu, et al. "Classification of extrachromosomal circular DNA with a focus on the role of extrachromosomal DNA (ecDNA) in tumor heterogeneity and progression." Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 2020.

[2] Peterson, Leif E., and Matthew A. Coleman. "Machine learning-based receiver operating characteristic (ROC) curves for crisp and fuzzy classification of DNA microarrays in cancer research." International Journal of Approximate Reasoning, vol.47, pp. 17-36, 2008.

[3] Winters-Hilt, Stephen, et al. "Highly accurate classification of Watson-Crick basepairs on termini of single DNA molecules." Biophysical Journal, vol.84, pp. 967-976, 2003.

[4] Feng, Ting-Cheng, Tzuu-Hseng S. Li, and Ping-Huan Kuo. "Variable coded hierarchical fuzzy classification model using DNA coding and evolutionary programming." Applied Mathematical Modelling vol.39, pp.23-24 (2015): 7401-7419.

[5] Atila, Ümit, et al. "Classification of DNA damages on segmented comet assay images using convolutional neural network." Computer methods and programs in biomedicine vol.186 (2020): 105192.

[6] Touati, R., et al. "New intraclass helitrons classification using DNA-image sequences and machine learning approaches." IRBM vol.42.3 (2021): pp.154-164.

[7] Liu, L., et al. "Targeted methylation sequencing of plasma cell-free DNA for cancer detection and classification." Annals of Oncology vol.29.6 (2018): pp.1445-1453.

[8] Liberatib, Sergio Bittantia Simone Garattia Diego. "From DNA Micro-Arrays to Disease Classification: an Unsupervised Clustering Approach." (2005).

[9] de Vries, Jard H., et al. "Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy." Forensic Science International: Genetics vol.356 (2022): 102625.

[10] Hu, Jun, et al. "TargetDBP: accurate DNA-binding protein prediction via sequence-based multi-view feature learning." IEEE/ACM transactions on computational biology and bioinformatics vol.17.4 (2019): pp.1419-1429.

[11] Wu, Lei, et al. "Profiling DNA mutation patterns by SERS fingerprinting for supervised cancer classification." Biosensors and Bioelectronics 165 (2020): 112392.

[12] Garro, Beatriz A., Katya Rodríguez, and Roberto A. Vázquez. "Classification of DNA microarrays using artificial neural networks and ABC algorithm." Applied Soft Computing 38 (2016): pp.548-560.

[13] Khan, Firoz, et al. "A digital DNA sequencing engine for ransomware detection using machine learning." IEEE Access 8 (2020): pp.119710-119719.

[14] Zhou, Jiyun, et al. "EL_LSTM: prediction of DNA-binding residue from protein sequence by combining long short-term memory and ensemble learning." IEEE/ACM transactions on computational biology and bioinformatics 17.1 (2018): pp.124-135.

[15] Rashed, Amr Ezz El-Din, et al. "Sequence Alignment Using Machine Learning-Based Needleman–Wunsch Algorithm." IEEE Access 9 (2021): pp.109522-109535.

[16] Zhang, Qinhu, Zhen Shen, and De-Shuang Huang. "Predicting in-vitro transcription factor binding sites using DNA sequence+ shape." IEEE/ACM transactions on computational biology and bioinformatics (2019).

[17] Aref-Eshghi, Erfan, et al. "Genomic DNA methylation signatures enable concurrent diagnosis and clinical genetic variant classification in neurodevelopmental syndromes." The American Journal of Human Genetics 102.1 (2018): pp.156-174.

[18] Fiannaca, Antonino, et al. "A k-mer-based barcode DNA classification methodology based on spectral representation and a neural gas network." Artificial intelligence in medicine 64.3 (2015): pp.173-184.

[19] Alghunaim, Sara, and Heyam H. Al-Baity. "On the scalability of machine-learning algorithms for breast cancer prediction in big data context." IEEE Access 7 (2019): pp.91535-91546.

[20] Liu, Wenbin, Dongdong Li, and Henry Han. "Manifold learning analysis for allele-skewed DNA modification SNPs for psychiatric disorders." IEEE Access 8 (2020): pp.33023-33038.

[21] Pavesi, Giulio, and Giorgio Valentini. "Classification of co-expressed genes from DNA regulatory regions." *Information Fusion* vol.10.3 (2009): pp.233-241.

[22] Kaur, Pinderpal, et al. "DNA damage protection: an excellent application of bioactive compounds." Bioresources and Bioprocessing vol.6.1 (2019): pp.1-11.

[23] Shi, Bingyang, et al. "Challenges in DNA delivery and recent advances in multifunctional polymeric DNA delivery systems." Biomacromolecules 18.8 (2017): pp.2231-2246.

[24] National Human Genome Research Institute: https://www.genome.gov/about-genomics/fact-sheets/Deoxyribonucleic-Acid-Fact-Sheet.

[25] Lindahl,T,Wood,RD, Reactome, Quality control by DNA repair: https://reactome.org/content/detail/R-HSA-73894.

[26] Ludovic Bourré, Crown bioscience: https://blog.crownbio.com/dna-damage-response.

[27] Lindström, Mikael S., Jiri Bartek, and Apolinar Maya-Mendoza. "p53 at the crossroad of DNA replication and ribosome biogenesis stress pathways." *Cell Death & Differentiation* (2022): 1-11.

[28] Lietz, Christopher E., Erik T. Newman, Andrew D. Kelly, David H. Xiang, Ziying Zhang, Caroline A. Luscko, Santiago A. Lozano-Calderon et al. "Genome-wide DNA methylation patterns reveal clinically relevant predictive and prognostic subtypes in human osteosarcoma." *Communications biology* 5, no. 1 (2022): 1-20.

[29] Valiente-Pallejà, A., Tortajada, J., Bulduk, B. K., Vilella, E., Garrabou, G., Muntané, G., & Martorell, L. (2022). Comprehensive summary of mitochondrial DNA alterations in the postmortem human brain: A systematic review. *EBioMedicine*, *76*, 103815.