Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 46 No. 5
May 2025

Analysis of DNA Sequence Classification Using Machine Learning Techniques

Madhushree Meghavath K B[1], Indrakumar K[1], Sharath H S[1] and Mohammed A.S Al-Mohamadi[1], Vighnesh H Y[2], Chiranjeevi [3]

[1]Department of Computer Science, Kuvempu University, Shimoga, Karnataka, INDIA
[2]Sri JCBM College, Chikkamagaluru, Karnataka, INDIA
[3]WebOccult Technologies, Ahmedabad, Gujarat, INDIA
[1]mmkb2401@gmail.com, [1]indk214@gmail.com, [1]sharathhs09@gmail.com, [1]almohmdy30@gmail.com,
[2]hyvighnesh93@gmail.com and [3]chiranjeevic1305@gmail.com

Abstract

Introduction: DNA serves as the fundamental genetic blueprint of life. Extracting meaningful patterns and information from DNA sequences is essential for advancements in genomics and comparative biology.

Objectives: This study aims to classify and distinguish DNA sequences from different species, human, chimpanzee, and dog using machine learning techniques to evaluate their effectiveness in genomic sequence analysis.

Methods: Three datasets consisting of DNA sequences from humans, chimpanzees, and dogs were used. Preprocessing included k-mer analysis and the CountVectorizer technique. Various machine learning algorithms, Naive Bayes, SVM, KNN, Decision Trees, and Random Forests were employed, alongside a Convolutional Neural Network (CNN) for deep learning-based classification. K-mers ranging from 5 to 8 were tested, with 6-mers yielding the best results.

Results: Naive Bayes achieved accuracies of 80% for human DNA, 87% for chimpanzee DNA, and 68% for dog DNA using 6-mers. CNN provided enhanced performance with 90.76% accuracy for human data, 83.64% for chimpanzee data, and 76.53% for dog data.

Conclusions: The use of 6-mers significantly improves classification accuracy across species. CNN models outperform traditional machine learning classifiers, demonstrating the potential of deep learning in genomic sequence analysis.

Keywords: DNA, Machine learning classifiers, K-mers, Countvectorizer.CNN.

1. Introduction

DNA, or “deoxyribonucleic acid, is a molecule that carries the genetic instructions” for the development, functioning, growth, and reproduction of all known living organisms and many viruses. It is often referred to as the "building block of life" due to its fundamental role in genetics. DNA consists of two long chains, known as strands, that are made up of smaller units called nucleotides. Each nucleotide contains a sugar molecule (deoxyribose), a phosphate group, and one of four nitrogenous bases: “adenine (A), thymine (T), cytosine (C), or guanine (G)”. The structure of DNA is often depicted as a double helix, where the two strands are twisted around each other in a spiral.

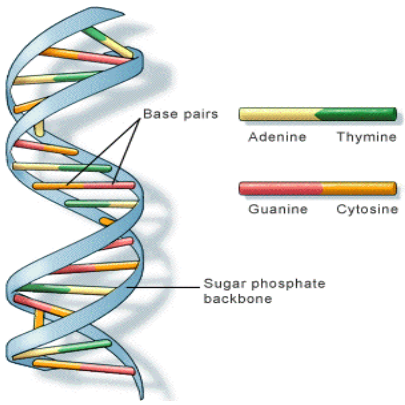


Figure 1: DNA Model

As shown in Figure 1, all living organisms have a genome, though the size and complexity of these genomes can vary significantly. In humans, the genetic material is organized across 23 chromosomes, and each genome contains over 6 billion DNA base pairs. Despite this immense size, most human genomes are remarkably similar to one another. The specific

arrangement of nitrogenous bases along the DNA molecules encodes the genetic blueprint that dictates an organism's traits and characteristics.

DNA strands are held together through base pairing, where adenine (A) pairs with thymine (T), and cytosine (C) pairs with guanine (G). This pairing mechanism is essential not only for copying DNA during replication but also for transcribing it into RNA. Figure 2 illustrates DNA sequencing a method used to determine the precise sequence of nucleotides in a DNA strand.

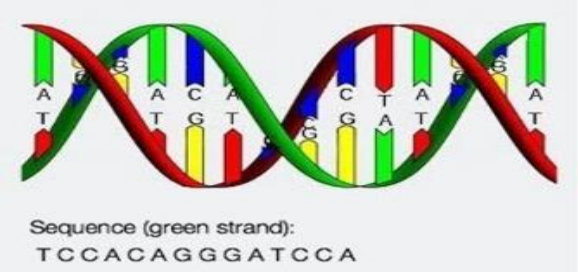


Figure 2: Sequence Strand

DNA, or Deoxyribonucleic Acid, serves as the hereditary material that holds the vital instructions necessary for an organism's growth, development, function, and reproduction. It acts as a fundamental guide for all biological processes in living beings. The sequence of nucleotide bases (adenine, thymine, cytosine, and guanine) within a DNA molecule determines the genetic characteristics of an organism. In the field of genomics, understanding DNA sequences and comparing them across species is critical to unlocking insights into evolutionary relationships, species identification, genetic diversity, and the molecular basis of diseases. DNA sequence classification, therefore, serves as an essential tool in bioinformatics, facilitating the identification of species or organisms based on their genetic material.

This study focuses on DNA sequence classification between three species humans, chimpanzees, and dogs using machine learning techniques, particularly Convolutional Neural Networks (CNNs). Each of these species represents different levels of genetic similarity and divergence, making them ideal for comparative analysis. Humans and chimpanzees share approximately 98-99% of their DNA, while dogs are significantly more genetically distinct. By comparing DNA sequences from these species, we aim to analyse how effectively machine learning algorithms can distinguish between organisms based on their genomic data. The ability to classify DNA sequences with high accuracy can not only aid in species identification but also provide a deeper understanding of evolutionary connections and genetic variations.

1.1 DNA Classification in Genomics

The process of DNA classification involves analyzing the nucleotide sequences of DNA and assigning them to predefined categories, typically based on species or genetic traits. In genomics, DNA sequence classification is widely used for tasks such as identifying genetic variations, understanding evolutionary relationships, detecting mutations, and diagnosing genetic disorders. It is a fundamental component of comparative genomics, which seeks to compare the genomes of different organisms to draw inferences about their biology, evolution, and development.

Classifying DNA sequences requires the use of advanced computational techniques, as DNA is made up of long strings of nucleotide bases that contain hidden patterns and features. Machine learning algorithms are particularly well-suited for this task as they can learn from large datasets and automatically identify meaningful patterns that may not be easily discernible through traditional analytical methods. By training models on labeled DNA sequences, machine learning algorithms can predict the species or characteristics of unseen DNA sequences, making them powerful tools for classification tasks in genomics.

1.2 Comparative Genomics: Human, Chimpanzee, and Dog DNA

Comparing the genomes of different species provides valuable insights into their evolutionary history and genetic makeup. Humans, chimpanzees, and dogs offer a compelling case study for DNA sequence classification due to their differing levels of genetic similarity. The human genome has been extensively studied and serves as a reference point for comparisons with other species. Chimpanzees, being the closest living relatives to humans, share a high degree of genetic similarity, with only about 1-2% of their DNA differing from that of humans. This close genetic relationship makes it challenging but insightful to classify DNA sequences from humans and chimpanzees, as it tests the machine learning models' ability to detect subtle differences between closely related species.

On the other hand, dogs are much more genetically distant from both humans and chimpanzees. The classification of dog DNA presents a different challenge, as the model needs to distinguish between organisms that share fewer common genetic features. The ability

to classify dog DNA accurately despite its divergence from human and chimpanzee DNA demonstrates the robustness and generalizability of the machine learning models used in this study.

1.3 Machine Learning and DNA Classification

Machine learning algorithms have revolutionized the field of bioinformatics by enabling the analysis of vast amounts of genomic data in ways that were previously impossible. Traditional methods of DNA analysis, such as sequence alignment and phylogenetic tree construction, are limited by their reliance on predefined patterns and rules. In contrast, machine learning algorithms can learn directly from the data, making them more flexible and powerful for tasks like DNA sequence classification.

A wide range of machine learning techniques has been applied to DNA classification, each with its strengths and limitations:

•Naive Bayes is a probabilistic classifier that works well with discrete data such as k-mer counts (short nucleotide sequences).

•Support Vector Machine (SVM) is a powerful classification algorithm that is widely used in bioinformatics for its ability to handle complex, high-dimensional datasets.

•K-Nearest Neighbors (KNN) is a simple yet effective method that classifies data points based on the closest training examples in the feature space.

•Decision Trees and Random Forests are tree-based models that split data based on feature values, making them highly interpretable and effective for non-linear relationships.

•Convolutional Neural Networks (CNNs) are deep learning models that have shown remarkable success in image classification and have been adapted for sequence data due to their ability to capture local patterns and hierarchies.

In this study, we focus on using a CNN model to classify DNA sequences from humans, chimpanzees, and dogs. CNNs have shown promise in bioinformatics applications, especially in tasks like protein structure prediction and DNA sequence analysis. The key advantage of CNNs lies in their convolutional layers, which can detect local patterns (such as k-mers) and learn hierarchical representations of the data. This makes them particularly well-suited for DNA sequence classification, where the model needs to capture both short-range and long-range dependencies in the sequence.

1.4 K-mer Analysis in DNA Classification

A central technique in DNA sequence classification is k-mer analysis, where DNA sequences are broken down into smaller overlapping subsequences (k-mers) of length k. Each k-mer represents a specific combination of nucleotide bases, and the frequency of occurrence of different k-mers provides valuable information about the genetic sequence. For example, the sequence "AGCT" could be split into 3-mers (subsequences of length 3) such as "AGC" and "GCT." By analyzing the distribution of k-mers within a DNA sequence, machine learning models can learn important features that distinguish between different species.

In this study, we use k-mer analysis along with the CountVectorizer method to preprocess the DNA sequences. The CountVectorizer method converts the DNA sequences into a matrix of k-mer counts, which serves as the input to the machine learning models. The length of the k-mers is an important hyperparameter that can affect the performance of the classifier. For this study, we experiment with different k-mer lengths and find that 6-mers yield the highest accuracy across all species. This suggests that 6-mers capture the optimal amount of information for distinguishing between human, chimpanzee, and dog DNA.

By training the CNN on DNA sequences from humans, chimpanzees, and dogs, we achieve high classification accuracy, with notable results of 90.76% for human data, 83.64% for chimpanzee data, and 76.53% for dog data. These results demonstrate the effectiveness of CNNs in capturing relevant features from DNA sequences and distinguishing between species based on their genetic material.

DNA sequence classification is an essential task in genomics, enabling researchers to identify species, understand evolutionary relationships, and detect genetic variations. The use of machine learning algorithms, particularly CNNs, has proven to be highly effective in this domain. By analyzing DNA sequences from humans, chimpanzees, and dogs, we have demonstrated the CNN's ability to accurately classify species based on their genetic data. The combination of k-mer analysis and CNNs allows for the extraction of meaningful features from DNA sequences, making

these methods valuable tools in the study of comparative genomics and bioinformatics.

Classification is a fundamental concept and task in the field of machine learning and data analysis. It serves several important purposes and has a wide range of applications across various domains. Classification in DNA sequencing is a crucial step in bioinformatics and genomics research. It plays a fundamental role in analyzing and interpreting the vast amount of genetic information obtained through DNA sequencing techniques. The machine learning algorithms employed for this classification task include MultiNomial Naive Bayes (NB), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees (DT), and Random Forest (RF).

•MultiNomial Naive Bayes (NB)

Naive Bayes, a probabilistic classification algorithm based on Bayes' theorem, simplifies computations by assuming independence among features, despite potential real-world dependencies. Its effectiveness in tasks like text classification and spam detection stems from calculating class probabilities using the simplified assumption. Though this assumption may not always hold in practical scenarios, Naive Bayes remains valuable for its efficiency in categorizing data

•Support Vector Machine (SVM)

Support Vector Machines (SVM) excel in classifying both linear and non-linear data by identifying the optimal hyperplane for maximal class separation. Particularly adept with high-dimensional and complex decision boundary scenarios, SVM employs a kernel trick to enhance separation by transforming data into higher-dimensional spaces.

•K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple and effective classification method that predicts based on the majority class among the k-nearest data points in the feature space, avoiding model training in lazy learning. The choice of k significantly influences predictive accuracy, and while KNN performs well across dataset sizes, computational efficiency becomes a concern with large datasets.

•Decision Tree Classifier (DT)

Decision Trees, a fundamental classification method, iteratively partitions data based on key attributes for simplicity and interpretability. Prone to overfitting, especially in deep structures, techniques like pruning and depth limitation are employed.

•Random Forest Classifier (RF)

Random Forest is an ensemble learning technique that enhances prediction accuracy by combining multiple decision trees. During training, it builds a forest of decision trees, with each tree trained on a random subset of the data with replacement. This approach helps reduce overfitting and enhances the model's ability to generalize to new data. Random Forest is robust to noisy data and is applicable for both classification and regression tasks, making it a versatile and powerful machine learning method.

•Convolutional Neural Network(CNN)

CNN, or Convolutional Neural Network, is a deep learning algorithm widely used for tasks involving grid-like data, such as image recognition or, in this case, DNA sequence analysis. CNNs are particularly effective at automatically learning features from raw data through a series of convolutional layers, pooling layers, and fully connected layers. In the context of DNA sequence classification, CNNs can capture complex patterns within the sequences, making them useful for improving accuracy in genomics tasks. By applying CNN to human, chimpanzee, and dog DNA sequences, we achieved higher classification accuracy compared to traditional machine learning models, particularly in capturing the local dependencies within the DNA k-mers.

Classification, a foundational concept in machine learning, is crucial for automating tasks and making informed decisions. Its applications span various fields, such as image recognition, medical diagnosis, and sentiment analysis in natural language processing. Classification plays a pivotal role in predictive modeling for fraud detection, recommendation systems, and customer churn prediction.

Genomic data analysis confronts challenges like managing massive volumes, addressing DNA structure complexity, and handling sequence variability. Selecting appropriate classification algorithms is vital, considering strengths and weaknesses, and deep genomics knowledge is essential for identifying relevant features. Dealing with imbalances, high dimensionality, and sequencing errors requires specialized preprocessing, and ensuring interpretability is crucial in scientific and medical applications. Ethical concerns include genomic data privacy, label accuracy,

and model transferability across datasets and species. Despite ethical considerations, overcoming classification challenges is essential for advancing DNA analysis's impact on genetics, healthcare, biotechnology, environmental research, agriculture, security, and personalized medicine.

This research aims to use machine learning for DNA sequence classification, comparing models based on accuracy across human, dog, and chimpanzee datasets. The primary goal is to identify the most consistently accurate model, determining its robustness for genetic classification tasks.

2. Objectives

The primary objective of this study is to develop and evaluate a robust framework for classifying DNA sequences from different species specifically, human, chimpanzee, and dog using a combination of traditional machine learning and deep learning techniques. The study aims to:

2.1 Compare the Performance of ML Algorithms: Assess and compare the effectiveness of various machine learning classifiers including Naive Bayes, Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees, and Random Forests in classifying genomic sequences from different species.

2.2 Apply k-mer Analysis for Feature Extraction: Investigate the impact of different k-mer sizes (5-mers, 6-mers, 7-mers, and 8-mers) on classification accuracy by converting DNA sequences into numerical feature vectors suitable for machine learning models.

2.3 Utilize CountVectorizer for DNA Representation: Employ the CountVectorizer technique for transforming raw DNA sequences into sparse matrix representations, facilitating model learning and comparison.

2.4 Implement a Deep Learning Approach (CNN): Design and train a Convolutional Neural Network (CNN) to automatically extract hierarchical features from DNA sequences, and benchmark its performance against traditional machine learning models.

2.5 Identify Optimal k-mer Size and Classifier: Determine which k-mer size yields the highest classification performance and which classifier (among both ML and DL methods) provides the best generalization across species datasets.

2.6 Evaluate Model Generalizability: Test the models across multiple species datasets (human, chimpanzee, dog) to evaluate consistency, robustness, and generalizability of the classification framework.

2.7 Lay the Groundwork for Future Integration: Provide insights that inform future research into hybrid models combining deep learning and classical ML methods, and their application in broader genomics contexts such as variant detection, gene annotation, and disease gene identification.

3. Literature Review

Ersoy Öz et al. [1] authored the book, "Support Vector Machines in DNA Sequencing Quality Control," explores SVM's application in classifying DNA sequencing data quality. Using SVMs, the study distinguishes between 'high' and 'low' quality sequences, achieving accurate labeling in 23 out of 24 chromatograms from the InSNP dataset. The novel approach combines feature extraction and SVMs, showcasing potential as a robust solution for automatic quality screening in DNA sequencing. Achieved an accuracy of 95.83%.

Teresita M. Porter et al. [2] book, "Rapid and Accurate Taxonomic Classification of Insect DNA Barcode Sequences," introduces a Naïve Bayesian classifier for swift and precise insect species identification using COI DNA sequences. Emphasizing the importance in ecological research, the book provides a valuable contribution to entomology and DNA barcoding, offering an efficient method for taxonomic classification.

Authored by Lailil Muflikhah et al. [3] the book "Prediction of Liver Cancer Based on DNA Sequence Using Ensemble Method" explores the link between chronic HBV infection and liver cancer. Through machine learning techniques, the study mitigates unbalanced data challenges, proposing an ensemble method with an 88.4% accuracy, 88.4% sensitivity, and 91.4% specificity in predicting liver cancer based on HBV DNA sequences.

Rodney T. Richardson et al. [4] investigate the performance of widely used DNA metabarcoding classification software in their paper "Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data." The study categorizes

classification methods, focusing on rdp Naïve Bayesian Classifier, rtax, and utax, and underscores the need for clear comparisons to improve the accuracy and reliability of DNA metabarcoding in ecological research. The research highlights challenges in selecting appropriate tools due to the diversity of available methods in DNA sequence classification.

Jiarong Guo, Ben Bolduc et al. [5] introduce "VirSorter2," an expert-guided tool for identifying DNA and RNA viruses. This sophisticated, freely accessible resource excels in taxonomic and functional diversity exploration of microbial communities, utilizing high-throughput gene marker and metagenomic sequencing technologies. VirSorter2's modular design facilitates updates, though it may be less sensitive for very short sequences; it stands out for differentiating eukaryotic genomes, plasmids, and viruses with high specificity, offering scalability for large datasets in virology and metagenomics. Achieved an accuracy of >80%.

Hemalatha Gunasekaran et al. [6] book, "Analysis of DNA Sequence Classification Using CNN and Hybrid Models," explores effective biomedical data analysis for virus identification. The study emphasizes the power of Convolutional Neural Networks (CNNs), specifically CNN, CNN-LSTM, and CNN-Bidirectional LSTM, showcasing k-mer encoding's superiority for accurate DNA sequence classification, achieving a notable 93.16% accuracy.

Frederick I. Archer et al. [7] investigate subspecies classification using machine learning, specifically Random Forests, applied to mitochondrial DNA (mtDNA) sequences. The study reveals insights into the impact of simulation parameters, such as migration and divergence time, on classification accuracy, ranging from 70% to 85%. The research sheds light on challenges and influential factors in accurately classifying genetic data for subspecies differentiation.

Maitena Tellaetxe-Abete et al. [8] introduce Ideafix, a decision tree-based algorithm refining variants in formalin-fixed and paraffin-embedded (FFPE) DNA sequencing data. Utilizing features like read pair orientation bias, genomic context, and variant allele frequency, Ideafix distinguishes deaminations from non-deaminations, outperforming existing tools with an accuracy of 96%.

Steven Salzberg et al. [9] introduce a gene-finding system using decision tree classifiers and dynamic programming, achieving a base-pair accuracy of 83% on human DNA sequences. The study emphasizes the system's preliminary stage, with plans for further refinement, including testing on a larger vertebrate DNA database and incorporating lookup information for enhanced performance.

Robert W. Jackson et al. [10] investigated twin zygosity determination in their study "Determination of Twin Zygosity: A Comparison of DNA with Various Questionnaire Indices." Using a subset of questions, a logistic regression achieved a high 91% correct classification rate for both monozygotic (MZ) and dizygotic (DZ) twins. The questionnaire demonstrated reliability and outperformed other methods in accuracy comparisons, suggesting its effectiveness as an alternative to DNA analysis when questionnaire validity is established within the twin cohort.

4. Methods

Figure 3 illustrates the methodology, which encompasses the components of dataset management, preprocessing in that two process applied k-mes and count vectorizer, and the final classification models.

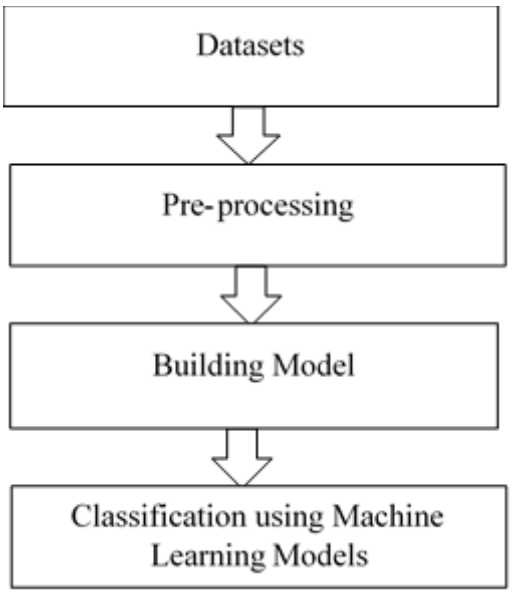


Figure 3: Block diagram of our Proposed Methodology

3.1 Datasets:

The dataset comprises text data consisting of DNA sequences from three distinct sources: humans, chimpanzees, and dogs. Each DNA sequence can be associated with anywhere from 0 to 6 class labels. Following this, a preprocessing phase is carried out on the dataset.

3.2 Preprocessing:

The dataset undergoes two important steps. First, k-mer counting is applied using a k-mer method to extract relevant patterns from the DNA sequences. Following this, the count vectorization process is employed to transform the k-mer counts into numerical features.

3.3 K-mer Counting:

In DNA sequence analysis, the method getKmers generates k-mers from sequences, using sizes 5, 6, 7, and 8, with 6-mers demonstrating higher accuracy. Applied to human, chimpanzee, and dog datasets, the k-mers are stored in new words columns, enabling subsequent text data processing for further analysis, such as text classification or machine learning tasks. The pattern matching process assesses the occurrence of specified patterns within the strings, providing counts for detected patterns.

3.4 Count Vectorizer:

Count vectorization, employed in genomics for DNA sequences, treats each unique 4-mer as a feature, creating a sparse matrix representing the count of each 4-mer in the sequences. This technique results in a "k-mer frequency matrix" or "count matrix" and is applied consistently across all three datasets.

3.5 Classifications:

Machine learning algorithms, namely Naive Bayes, SVM (Linear), KNN, Decision Tree (Gini), and Random Forest, are employed on human, chimpanzee, and dog datasets. The datasets are split into training (80%) and testing (20%), with allocations: 3,504 training, 876 testing for humans; 1,345 training, 337 testing for chimpanzees; and 656 training, 164 testing for dogs. The classifiers are trained on these subsets and evaluated on the corresponding testing sets.

•Naive Bayes Classifier

The Naive Bayes classifier in DNA sequence analysis involves preprocessing with the kmers method, followed by count vectorization to convert sequences into numerical format. Configured with Laplace smoothing (alpha=0.1), it's trained on labelled datasets for human, chimpanzee, and dog DNA sequences. During classification, it calculates probabilities for unlabelled sequences, achieving accurate categorization into the three classes. Evaluation metrics are applied to assess its performance.

•SVM Classifier

In the SVM classifier, datasets representing human, chimpanzee, and dog DNA sequences undergo preprocessing with kmers and count vectorization. The linear kernel with a gamma value of 1 is employed for training, suitable for high-dimensional data like DNA sequences. The trained model is evaluated using appropriate metrics, ensuring accurate classification into human, chimpanzee, or dog categories.

•KNN Classifier

The KNN classifier involves preprocessing with the kmers method and count vectorization for human, chimpanzee, and dog DNA sequences. The KNeighbors Classifier is configured with neighbors=5 and p=2, utilizing the Euclidean distance measure. The model is trained on the datasets, and evaluation metrics are applied to gauge its performance in accurate categorization.

•Decision Tree Classifier

For the Decision Tree classifier, preprocessing utilizes the kmer method and count vectorization for human, chimpanzee, and dog DNA sequences. The Decision Tree Classifier is configured with the gini criterion, assessing impurity for optimal node splits during training. Evaluation metrics are applied to assess the model's performance in classifying DNA sequences into the respective categories.

•Random Forest

In Random Forest classification, preprocessing involves kmers and count vectorization for human, chimpanzee, and dog DNA sequences. The RandomForest Classifier is configured with 100 decision trees and a fixed random state. The model is trained and evaluated using appropriate metrics for robust classification into human, chimpanzee, or dog categories.

•Convolutional Neural Network(CNN)

This CNN model is structured to classify sequence data [Figure 4], such as DNA sequences, using several layers that progressively extract and process features from the input data. The model starts with a Sequential structure, which allows stacking layers sequentially where each layer has a direct input and output relationship.

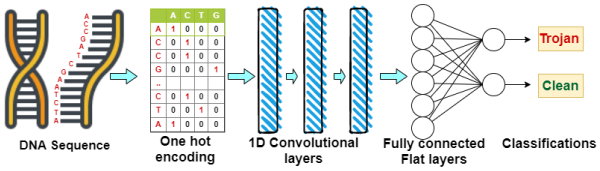


Figure 4: Architecture of CNN

The first layer is a 1D Convolutional Layer (Conv1D). It applies convolution operations on the input data using 64 different filters (or feature detectors), each of which is of length 3. These filters move across the input DNA sequences to extract important features or patterns. The ReLU (Rectified Linear Unit) activation function introduces non-linearity, enabling the model to learn complex patterns within the sequences. The input data is shaped as `(maxlen, 1)`, where `maxlen` is the length of the sequence, and `1` represents that we are working with a single channel (such as nucleotide information in DNA sequences).

After the convolutional layer, a MaxPooling1D layer is applied. This layer performs a downsampling operation, reducing the dimensionality of the feature maps produced by the convolutional layer. It pools the maximum value within a window of size 2, effectively retaining the most important features while reducing the number of parameters and computations. This also helps in reducing overfitting by simplifying the model's representation of the data.

Next, a Flatten layer is used, which reshapes the pooled feature maps into a 1-dimensional vector. This is necessary because the upcoming dense (fully connected) layers require a flattened input to perform classification tasks. Once the data is flattened, it is passed into a Dense layer with 100 neurons. This fully connected layer applies learned weights to the features extracted by the convolutional layers and introduces a non-linear transformation using the ReLU activation function. The layer captures complex feature interactions and relationships in the data.

To prevent overfitting, a Dropout layer with a rate of 0.5 is introduced. During each iteration of training, this layer randomly drops out (sets to zero) 50% of the neurons in the dense layer, ensuring that the model does not become overly reliant on specific neurons and is better able to generalize to unseen data.

Finally, the output is processed by another Dense layer with a number of neurons equal to the number of classes (DNA classifications), using the softmax activation function. Softmax converts the raw output values into a probability distribution across the classes, which makes it possible to classify the input into one of the categories.

Once the model is defined, it is compiled using the Adam optimizer, which is an adaptive learning rate optimization algorithm. Adam is chosen for its efficiency and ability to handle sparse gradients. The categorical crossentropy loss function is used because the classification task involves multiple classes, and the model's performance is evaluated using accuracy as the metric.

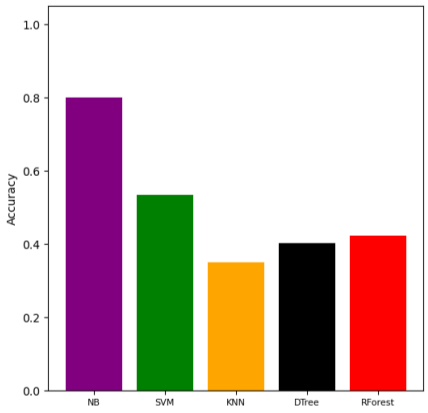
The model is then trained using the fit function. The training data (padded DNA sequences) and corresponding one-hot encoded labels are provided to the model. Training runs for 10 epochs, where in each epoch, the model processes the entire dataset, but in batches of 16 samples at a time. Additionally, 20% of the training data is set aside as a validation set, allowing the model to monitor its performance on unseen data during training.

This CNN model is particularly well-suited for tasks like DNA sequence classification, where local patterns within the sequences (k-mers) are important. The convolutional layers automatically learn these patterns and the combination of pooling, dense layers, and dropout ensures the model is both powerful and generalizable.

5. Results and Discussion

This study employed the k-mer method using values of 5, 6, 7, and 8. Among these, the utilization of 6 demonstrated superior accuracy, leading to a detailed discussion of the outcomes generated by the k-mer value of 6. In Figure 5 diverse classifiers, including Naive Bayes, SVM, KNN, Decision Trees, Random Forests and CNN were evaluated on the human DNA dataset. Following thorough experimentation, Naive Bayes consistently outperformed, proving its accuracy and reliability in predicting patterns within human DNA sequences.

Figure 5: Comparison of Classifiers for Human DNA Sequence



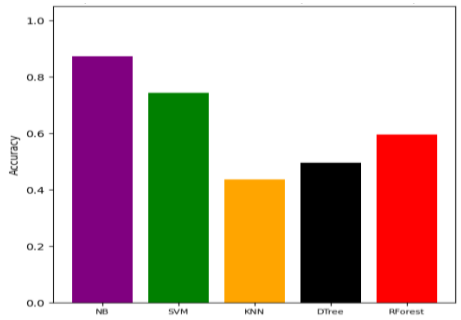


Figure 6: Comparison of Classifiers for Chimpanzee DNA Sequence

Within Figure 6, which illustrates the comparison of classifiers for the analysis of chimpanzee DNA sequences, it was determined that among the machine learning algorithms employed, Naive Bayes yielded the most favourable results. The similarity between chimpanzee DNA sequences and human DNA sequences was found to be significant. Consequently, for both of these datasets, Naive Bayes was identified as the most effective classifier.

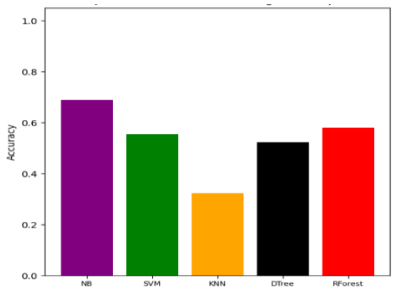


Figure 7: Comparison of Classifiers for Dog DNA Sequence

In the dog DNA sequence comparison (Figure 7), Naive Bayes emerges as the top performer, attributed to its efficiency and effectiveness in handling text-like data. However, classifier selection depends on the dataset and problem, necessitating consideration of diverse factors for optimal choice.

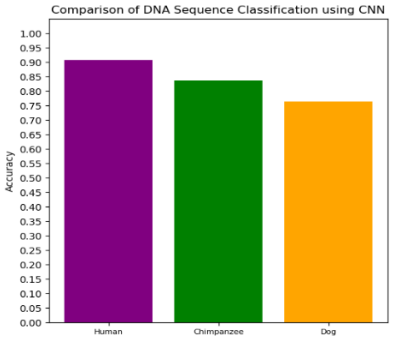


Figure 8: Comparison of DNA Sequence Classification using CNN

Figure 8 illustrates the performance of the CNN classifier in classifying DNA sequences from three different species: human, chimpanzee, and dog. The CNN model demonstrates its effectiveness in capturing complex patterns within DNA sequences, achieving high accuracy across all datasets. Specifically, the model achieves an accuracy of 90.76% for human DNA, indicating its strong ability to correctly identify human-specific genetic sequences. For chimpanzee DNA, the model attains an accuracy of 83.64%, highlighting its proficiency in distinguishing closely related species with subtle genetic differences. In the case of dog DNA, the CNN classifier achieves a notable accuracy of 76.53%, despite the greater genetic divergence from the human and chimpanzee datasets.

These results underscore the CNN's capability to generalize across varying DNA sequences, leveraging its convolutional layers to extract meaningful features from the input data. The variation in accuracy across the species can be attributed to differences in the genetic complexity and similarity between the datasets. Overall, the use of CNN as a classifier significantly enhances the performance of DNA sequence classification, especially when compared to traditional machine learning approaches, by providing more precise and reliable results for each species.

In the table 1, illustrates the accuracy of all models, which helps to understand how effectively each of these classifiers can distinguish between DNA sequences linked to humans, chimpanzees, and dogs. Summarizing the performance of these models in classifying DNA sequences for these species, the following insights:

Table 1: Accuracy of all models.

Model	Human	Chimpanzee	Dog
Naïve Bayes	72%	81%	64%
SVM	57%	81%	68%
KNN	34%	41%	34%
DTree	37%	41%	52%
RForest	46%	65%	60%
CNN	90.76%	83.64%	76.53%

From table 1, the classification of Humans, Naive Bayes achieved the leading accuracy of 80%, followed by SVM at 53%, KNN at 35%, Decision Tree at 40%, and Random Forest at 42%.In Chimpanzee classification, Naive Bayes displayed the highest accuracy of 87%, followed by SVM at 74%, KNN at 43%, Decision Tree at 49%, and Random Forest at 59%.Regarding Dog classification, Naive Bayes

secured an accuracy of 68%, SVM achieved 55%, KNN had 32%, Decision Tree reached 52%, and Random Forest attained 57%.

Other performance metrics are:

•For Naive Bayes Classifier:

The Naive Bayes classifier demonstrates superior accuracy compared to other methods across all three datasets. Consequently, the confusion matrix is specifically showcased for this classifier.

Above Figure 9 explains that on the Human Test DNA Sequence, it demonstrates accurate positive classification with a accuracy of 80.1%, precision of 82.3%, effectively capturing actual positive instances with a recall of 80.1%. The F1 Score of 80.2% reflects overall good performance.

Predicted	0	1	2	3	4	5	6
Actual							
0	85	0	0	0	1	0	16
1	5	85	0	2	0	0	14
2	1	0	63	3	1	0	10
3	14	0	3	82	0	0	26
4	5	2	1	9	104	0	28
5	2	2	0	4	0	37	6
6	2	0	2	11	4	0	246

accuracy_nb_human = 0.801
precision = 0.823
recall = 0.801
f1 = 0.802

Figure 9: Confusion matrix for human data using Naïve Bayes classifier

Figure 10 says that, For the Chimpanzee Test DNA Sequence the model showcases exceptional accuracy at 87.2%, precision at 88.9% and strong recall at 87.2%, resulting in a robust F1 Score of 87.2%. It exhibits high accuracy and reliability in classifying positive instances.

Predicted	0	1	2	3	4	5	6
Actual							
0	28	0	0	0	0	0	0
1	2	32	0	0	0	0	5
2	0	0	21	2	0	0	4
3	2	0	0	33	1	0	7
4	3	0	0	3	38	0	8
5	1	0	0	0	0	26	2
6	1	0	0	2	0	0	116

accuracy_nb_chimp = 0.872
precision = 0.889
recall = 0.872
f1 = 0.872

Figure 10: Confusion matrix for chimpanzee data using Naïve Bayes classifier

Within Figure 11, On the Dog Test DNA Sequence, the model achieves accuracy at 68.9%, decent precision at 78.1%, but its recall is lower at 68.9%, leading to an F1 Score of 67.3%. While accuracy is acceptable.

Predicted	0	1	2	3	4	5	6
Actual							
0	21	0	0	0	0	0	6
1	3	10	0	0	0	0	6
2	1	0	10	0	0	0	3
3	1	0	0	8	0	0	7
4	3	0	0	2	8	0	10
5	2	0	0	1	0	5	5
6	0	0	0	1	0	0	51

accuracy_nb_dog = 0.689
precision = 0.781
recall = 0.689
f1 = 0.673

Figure 11: Confusion matrix for dog using Naïve Bayes classifier

•For SVM Classifier:

For human DNA sequences, precision is 77.6%, recall is 53.5%, and the F1 score is 51.0%. On chimpanzee DNA sequences, it achieves precision of 82.9%, recall of 74.5%, and an impressive F1 score of 73.8%. However, on dog DNA sequences, the model's precision is 68.0%, recall is 55.5%, and the F1 score is 52.3%.

•For KNN Classifier:

For human DNA sequences, it achieves a precision of 67.9%, recall of 35.2%, and an F1 score of 24.8%. On chimpanzee DNA sequences, the model attains a precision of 70.0%, recall of 43.6%, and an F1 score of 33.2%. However, on dog DNA sequences, precision is 40.6%, recall is 32.3%, and the F1 score is 22.0%.

•For Decision Tree Classifier:

When dealing with human DNA sequences, it achieves a precision rate of 42.4%, a recall rate of 40.2%, and an F1 score of 37.7%. On the other hand, when applied to chimpanzee DNA sequences, the model demonstrates a precision rate of 50.8%, a recall rate of 49.6%, and an F1 score of 48.4%. Similarly, for dog DNA sequences, the model exhibits a precision rate of 52.4%, a recall rate of 52.4%, and an F1 score of 51.5%.

•For Random Forest Classifier:

When applied to human DNA sequences, it achieves a precision rate of 60.3%, a recall rate of 42.4%, and an F1 score of 34.0%. For chimpanzee DNA sequences, the model showcases a precision rate of 70.7%, a recall rate of 59.6%, and an F1 score of 55.9%. Similarly, in the case of dog DNA sequences, the model displays a precision rate of 70.1%, a recall rate of 57.9%, and an F1 score of 55.5%.

•For CNN Classifier:

Using a Convolutional Neural Network (CNN) for DNA sequence classification, we achieved distinct accuracies across three species: 90.76% for human data, 83.64% for chimpanzee data, and 76.53% for dog data. The CNN model demonstrated its ability to effectively learn and identify patterns within the genetic sequences of each species, with human DNA classification yielding the highest accuracy. This suggests that the CNN was particularly successful at recognizing specific features unique to human genetic data.

For chimpanzee data, the model performed well, reflecting the genetic closeness between humans and chimpanzees. Although slightly lower than for human DNA, the 83.64% accuracy shows the model's capacity to differentiate between these two closely related species. When applied to dog DNA sequences, the CNN achieved a lower but still significant accuracy of 76.53%, which can be attributed to the greater genetic divergence between dogs and the other two species. Despite this, the model was still able to extract meaningful patterns from the dog DNA sequences to classify them with reasonable accuracy. Overall, the CNN's performance across these datasets highlights its strength in identifying local patterns in genetic sequences and demonstrates its effectiveness in distinguishing between species with varying degrees of genetic similarity.

•Graph Comparision:

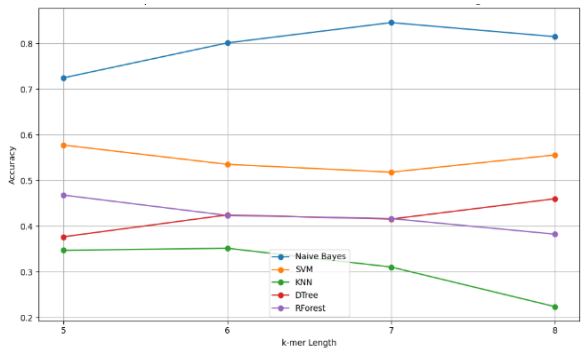


Figure 12: Comparision of model accuracies for human data at different k-mer lengths

In (Figure 12) the given graph illustrating the comparison of model accuracies for human data across various k-mer lengths, it's observed that Naive Bayes consistently outperforms followed by other classification models, such as SVM, Random Forest, Decision Tree, then KNN across all k-mer lengths (5, 6, 7, and 8), showcasing higher accuracy rates.

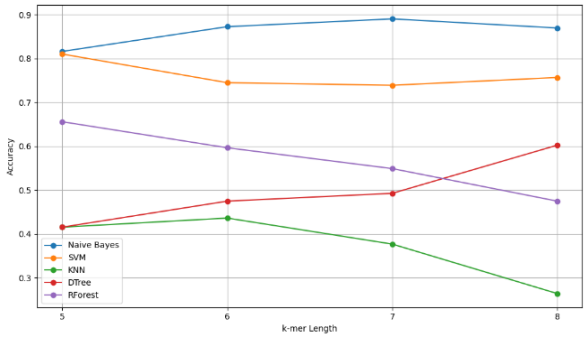


Figure 13: Comparision of model accuracies for chimpanzee data at different k-mer lengths

The graph in (Figure 13) displays the contrast in model accuracies for chimpazee data across different k-mer lengths. It's evident that Naive Bayes consistently outperforms followed by alternative classification models like SVM , Random Forest, Decision Tree, then KNN at k-mer lengths 5, 6, 7, and 8, demonstrating superior accuracy rates.

In (Figure 14), the graph exhibits the variation in model accuracies for dog data across various k-mer lengths. Naive Bayes emerges as the top performer, followed by alternative classifiers such as SVM, Random Forest, Decision Tree, and then KNN, at k-mer lengths 5, 6, 7, and 8, showcasing superior accuracy rates.

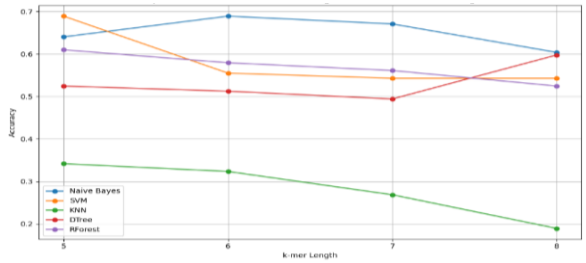
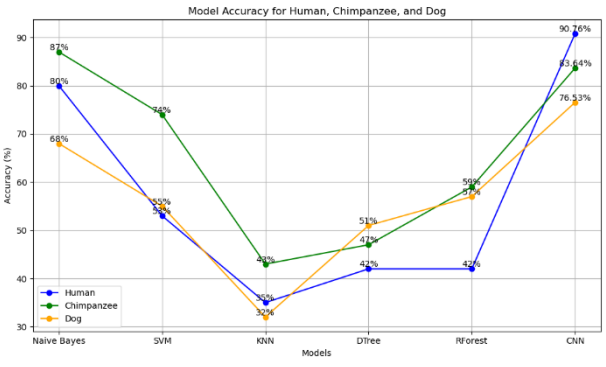


Figure 14: Comparision of model accuracies for dog data at different k-mer lengths



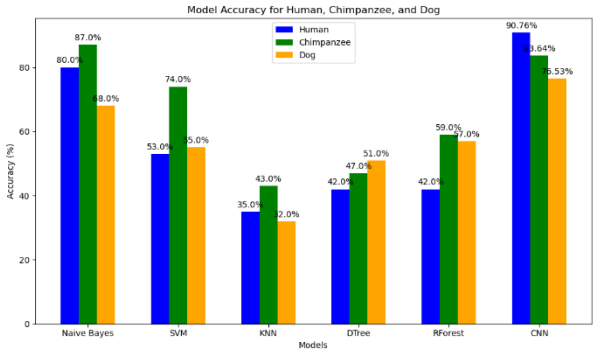


Figure 15: Bar Chart Comparison of model accuracies for human, chimpanzee and dog data.

In Figure 15, a bar chart is presented to visually compare the classification accuracies of the CNN model across three species: human, chimpanzee, and dog. The height of each bar corresponds to the accuracy achieved for each dataset, with human data reaching the highest accuracy at 90.76%, followed by chimpanzee data at 83.64%, and dog data at 76.53%. This bar chart offers a clear and straightforward way to compare the performance of the model on different datasets, highlighting the differences in classification accuracy between species. The chart effectively emphasizes the CNN's proficiency in handling human and chimpanzee data compared to dog data, showcasing how genetic similarities and divergences impact the model's accuracy.

In Figure 16, a line graph is used to represent the same classification accuracies of the CNN model for human, chimpanzee, and dog data. The line graph provides a continuous visual connection between the accuracy values for the three species, illustrating the trend in model performance. As the line progresses, it shows the highest point for human data, a slightly lower point for chimpanzee data, and a further drop for dog data. This format of visualization highlights the gradual decline in accuracy as the genetic divergence between species increases, offering a clearer view of the pattern in model performance compared to the bar chart.

Both figures serve the purpose of visually comparing the model's performance on different datasets, with the bar chart offering a categorical comparison and the line graph showing the trend in accuracy changes across species.

Figure 16: Line Graph Comparison of model accuracies for human, chimpanzee and dog data.

6. Conclusion

In conclusion, DNA sequencing is pivotal in genetics and genomics, utilizing ML methods such as Naive Bayes, SVM, KNN, Decision Trees, and Random Forests to discern genes, including disease-causing ones. Notably, among k-mers 5, 6, 7, and 8, k-mer 6 stands out, where Naive Bayes consistently achieves peak accuracy 80% for humans, 87% for chimpanzees, and 68% for dogs. Classifier choice significantly impacts accuracy, with Naive Bayes excelling for all 3 datasets of humans, chimpanzee and dogs. The integration of machine learning in genomics enhances classification accuracy, deepening our insights into genetic functions and disease research. DNA sequencing, vital for unraveling life's mysteries, underscores the potential of ML in advancing genomics knowledge.

Future work in DNA sequencing entails the integration of diverse machine learning techniques, encompassing both deep learning and traditional ML methods. This includes hybrid models that combine deep learning's feature extraction capabilities with traditional ML classifiers. Researchers will explore transfer learning, interpretability, ensemble methods, and data augmentation to enhance classification and variant calling accuracy. Integrating multi-model data, ensuring scalability, addressing ethical considerations, and facilitating clinical implementation are key aspects of future genomics research.

Refrences

[1] Ersoy Öz and Hüseyin Kaya, "Support Vector Machines in DNA Sequencing Quality Control",2013

[2] Teresita M. Porter, Joel F. Gibson, Shadi Shokralla, Donald J. Baird, G. Brian Golding, and Mehrdad Hajibabaei, "Rapid and Accurate Taxonomic Classification of Insect (Class Insecta) Cytochrome c Oxidase Subunit 1 (COI) DNA Barcode Sequences Using a Naïve Bayesian Classifier",2014

[3] Lailil Muflikhah, Nashi Widodo, Wayan Firdaus Mahmudy, and Solimun ,"Prediction of Liver Cancer Based on DNA Sequence Using Ensemble Method", Universitas Brawijaya Malang, East Java, Indonesia,2020

[4] Rodney T. Richardson, Johan Bengtsson-Palme, Reed M. Johnson,"Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data",2016

[5] Jiarong Guo, Ben Bolduc, Ahmed A. Zayed, Arvind Varsani, Guillermo Dominguez-Huerta, Tom O.

Delmont, Akbar Adjie Pratama, M. Consuelo Gazitúa, Dean Vik, Matthew B. Sullivan, and Simon Roux, “VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses”, 2021

[6] Hemalatha Gunasekaran, K. Ramalakshmi, A. Rex Macedo Arokiaraj, S. Deepa Kanmani, Chandran Venkatesan, and C. Suresh Gnana Dhas ,"Analysis of DNA Sequence Classification Using CNN and Hybrid Models", 2021

[7] Frederick I. Archer, Karen K. Martien, and Barbara L. Taylor, " Diagnosability of mtDNA with Random Forests: Using sequence data to delimit subspecies", 2017

[8] Maitena Tellaetxe-Abete, Borja Calvo, and Charles Lawrie ,"Ideafix: a decision tree-based method for the refinement of variants in FFPE DNA sequencing data", 2021

[9] Steven Salzberg, Xin Chen, John Henderson, and Kenneth Fasman ,"Finding Genes in DNA using Decision Trees and Dynamic Programming", Department of Computer Science and Division of Biomedical Information Sciences Johns Hopkins University Baltimore,1996

[10] Robert W. Jackson, Harold Snieder, Harry Davis, and Frank A. Treiber ,“Determination of Twin Zygosity: A Comparison of DNA with Various Questionnaire Indices”, Cambridge University ,2012

[11] Nayak V., Mishra J., Naik M., Swapnarekha B., Cengiz H., Shanmuganathan K. An impact study of COVID-19 on six different industries: automobile, energy and power, agriculture, education, traveland tourism and consumer electronics. Expert Systems. 2021:1–32. doi: 10.1111/exsy.12677.

[12] Shadab S., Alam Khan M. T., Neezi N. A., Adilina S., Shatabda S. DeepDBP: deep neural networks for identification of DNA-binding proteins. Informatics in Medicine Unlocked. 2020;19, article 100318.

[13] Benson D. A., Karsch-Mizrachi I., Lipman D. J., Ostell J., Sayers E. W. GenBank. Nucleic Acids Research. 2010;38(Supplement 1):46–51. doi: 10.1093/nar/gkp1024.

[14] Momenzadeh M., Sehhati M., Rabbani H. Using hidden Markov model to predict recurrence of breast cancer based on sequential patterns in gene expression profiles. Journal of Biomedical Informatics. 2020;111, article 103570 doi: 10.1016/j.jbi.2020.103570.

[15] Solis-Reyes S., Avino M., Poon A. F. Y., Kari L. An Open-Source k-mer Based Machine Learning Tool for Fast and Accurate Subtyping of HIV-1 Genomes. bioRxiv; 2018.

[16] Karagöz M. A., Nalbantoglu O. U. Taxonomic classification of metagenomic sequences from Relative Abundance Index profiles using deep learning. Biomedical Signal Processing and Control. 2021;67, article 102539 doi: 10.1016/j.bspc.2021.102539.

[17] Deorowicz S. FQSqueezer: k-mer-based compression of sequencing data. Scientific Reports. 2020;10(1):578–579. doi: 10.1038/s41598-020-57452-6.

[18] Suriya M., Chandran V., Sumithra M. G. Enhanced deep convolutional neural network for malarial parasite classification. International Journal of Computers and Applications. 2019:1–10.

[19] Jang B., Kim M., Harerimana G., Kang S. U., Kim J. W. Bi-LSTM model to increase accuracy in text classification: combining word2vec CNN and attention mechanism. Applied Sciences. 2020;10(17):p. 5841. doi: 10.3390/app10175841.

[20] Zhang X., Beinke B., Al Kindhi B., Wiering M. Comparing machine learning algorithms with or without feature extraction for DNA classification. 2020.

[21] Do D. T., Le N. Q. K. Using extreme gradient boosting to identify origin of replication in Saccharomyces cerevisiae via hybrid features. Genomics. 2020;112(3):2445–2451. doi: 10.1016/j.ygeno.2020.01.017.

[22] Xu H., Jia P., Zhao Z. Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. Briefings in Bioinformatics. 2021;22(3):1–13. doi: 10.1093/bib/bbaa099.

[23] Nugent C. M., Adamowicz S. J. Alignment-free classification of COI DNA barcode data with the Python package Alfie. Metabarcoding and Metagenomics. 2020;4:81–89. doi: 10.3897/mbmg.4.55815.

[24] Remita A. M., Diallo A. B. Statistical linear models in virus genomic alignment-free classification: application to hepatitis C viruses. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); November 2019; San Diego, CA, USA.

[25] Lopez-Rincon A., Tonda A., Mendoza-Maldonado L., et al. Classification and specific primer design for

accurate detection of SARS- CoV-2 using deep learning. Scientific Reports. 2021;11(1):1–11. doi: 10.1038/s41598-020-80363-5.

[26] Arruda M. M., De Assis F. M., De Souza T. A. Is BCH code useful to DNA classification as an alignment-free method? IEEE Access. 2021;9:68552–68560.

[27] Maalik S. W. I., Ananta S. K. W. Comparation analysis of ensemble technique with boosting (Xgboost) and bagging (Randomforest) for classify splice junction DNA sequence category. Jurnal Penelitian Pos dan Informatika.

[28] Hussain F., Saeed U., Muhammad G., Islam N., Sheikh G. S. Classifying cancer patients based on DNA sequences using machine learning. Journal of Medical Imaging and Health Informatics. 2019;9(3):436–443. doi: 10.1166/jmihi.2019.2602.

[29] Ben Nasr F., Oueslati A. E. CNN for human exons and introns classification. 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD); March 2021; Monastir, Tunisia. pp. 249–254.

[30] Al-Ajlan A., El Allali A. CNN-MGP: convolutional neural networks for metagenomics gene prediction. Interdisciplinary Sciences: Computational Life Sciences. 2019;11(4):628–635.

[31] Kassim N. A., Abdullah A. Classification of DNA sequences using convolutional neural network approach. UTM Computing Proceedings Innovations in Computing Technology and Applications. 2017;2:1–6.

[32] Morales J. A., Saldaña R., Santana-Castolo M. H., et al. Deep learning for the classification of genomic signals. Mathematical Problems in Engineering. 2020;2020:9.