Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

A Novel DNA Classification Experiment: Spatial Transcriptomics Analysis for Human Monkeypox DNA-Motifs with Kolmogorov–Arnold Networks

Selçuk YAZAR[1*]

[1] Kırklareli University, Software Engineering Department, selcukyazar@klu.edu.tr, Orcid No: 0000-0001-6567-4995

ARTICLE INFO

Article history:

Received 22 August 2024
Received in revised form 25 September 2024
Accepted 8 October 2024
Available online 23 December 2024

Keywords:

Spatial Transcriptomics, Kolmogorov-Arnold Networks, Deep Learning, Gene Expression Patterns, Regional Classification

Doi: 10.24012/dumf.1537079

* Corresponding author

ABSTRACT

Spatial Transcriptomics(ST) has emerged as a powerful tool for understanding gene expression patterns across different regions of a tissue or organism. It is crucial for disease research and developing new therapies. It allows for the measurement of gene expression across specific, localized areas of a tissue slide, though it does so with limited throughput. Yet, the data produced by ST technologies are characteristically noisy, high-dimensional, sparse, and multi-modal, encompassing elements like histological images and count matrices. Existing methods for analyzing ST data, which often rely on traditional statistical or machine learning techniques, have proven inadequate in many cases due to challenges like scale, multi-modality, and the inherent limitations of spatially-resolved data, including spatial resolution, sensitivity, and gene coverage. To address these specific challenges, researchers have turned to deep learning-based models. In this study, we present a novel approach to transcriptomics analysis using Kolmogorov-Arnold Networks (KANs), a state-of-the-art deep learning model to predict regional origin of monkeypox transcriptomic sample. By leveraging the ability of KANs to learn and represent complex, non-linear functions, we aim to uncover intricate spatial patterns of gene expression and gain insights into the underlying biological processes. Study's analysis focuses on two distinct regions, America and Asia, and employs a KAN-based classifier. The results demonstrate the promising performance of KANs in this context, with a precision of 0.45 and a recall of 0.93 for the America region, indicating a strong ability to correctly identify samples from this region. Findings indicate that predicting the regional transcriptome of monkeypox from DNA motifs could facilitate image-based screening for phylogenetic analyses.

Introduction

Spatial transcriptomics (ST) represents a significant advancement over traditional gene expression analysis methods by incorporating the geographical location of DNA within tissue sections, thereby preserving the spatial context of gene activity. Traditional methods, such as single-cell RNA sequencing (scRNAseq), lack the ability to maintain this spatial information, which is crucial for understanding the complex interactions and heterogeneity within tissues [1]. ST technologies, by contrast, enable the profiling of gene expression at a single-cell resolution while retaining the cellular compositions within a tissue, offering insights into cellular interactions that were previously unattainable [2]. The data produced by conventional ST technologies exhibit inherent characteristics such as noise, high dimensionality, sparsity, and multimodality, necessitating the utilization of specialized computational tools like machine learning (ML) for precise and robust analysis [2]. The integration of machine learning (ML) techniques, has further differentiated ST from traditional methods. The development of tools like SPADE for identifying spatially variable genes and PERSIST for optimizing gene panels for ST studies exemplifies the tailored approaches being developed to leverage the unique aspects of ST data [3, 4]. These tools, which are based on machine learning models, offer superior performance in detecting spatially relevant gene expression patterns and selecting informative gene targets, respectively, compared to methods used in traditional gene expression analysis [5], [6]. Additionally, the application of ST, combined with graph-based machine learning methods, has been demonstrated in research on glioblastoma multiforme, uncovering spatially restricted tumor niches and signaling networks relevant to patient survival. This highlights the potential of ST combined with ML to contribute to the development of new therapeutic strategies by providing a more nuanced understanding of disease pathology at the spatial level. ST, enhanced by machine learning techniques, offers a more comprehensive and nuanced understanding of gene expression by preserving and analyzing the spatial context of tissues, a capability that traditional methods lack. Yet, machine learning frameworks have shown to be sub-optimal for analyzing the complex, noisy, and high-dimensional data generated by ST due to challenges such as spatial resolution, sensitivity, and gene coverage [7]. Deep learning (DL)-based models, however, are being developed to address these ST-specific challenges, including alignment, spatial reconstruction, and spatial clustering, showcasing the potential for transformational applications

in analyzing spatially resolved transcriptomics data [8]. These approaches not only advance our fundamental understanding of biological processes but also open new avenues for disease research and therapeutic development.

The historical process of human monkeypox disease traces back to its first identification in the Democratic Republic of Congo (DRC) in 1970, following the discovery of the monkeypox virus in monkeys in 1958 [9]. Initially endemic to Central and Western Africa, the disease is caused by the monkeypox virus (MPXV), a zoonotic orthopoxvirus sharing clinical similarities with smallpox but distinguishable by symptoms such as lymphadenopathy [10]. Over the years, the disease has seen a gradual increase in incidence, with a notable shift in the median age of affected individuals from children to young adults and a variation in fatality rates between different clades of the virus. Human monkeypox, caused by Monkeypox virus, has historical significance post smallpox eradication. The recent 2022 outbreak, with global spread and human-to-human transmission, highlights its current threat level.

The classification of human monkeypox disease using deep learning techniques encompasses a variety of approaches aimed at enhancing early detection, diagnosis, and understanding of the disease's spread. These methods leverage Convolutional Neural Networks (CNNs), transfer learning, ensemble learning, and feature fusion techniques to analyze skin lesion images and predict monkeypox outbreaks. One primary method involves the development of diagnostic models using Generalization and Regularization-based Transfer Learning approaches (GRA-TLA) for binary and multiclass classification of monkeypox, demonstrating the potential of machine learning in distinguishing between infected and non-infected individuals with high accuracy [11]. Similarly, the construction of a Computer-Aided Diagnosis (CAD) tool, "Monkey-CAD," utilizes features extracted from multiple CNNs, employing Discrete Wavelet Transform (DWT) for feature fusion and entropy-based feature selection to enhance classification performance [12]. The application of machine learning and image processing methods, including data augmentation and transfer learning strategies across various deep learning models, has been pivotal in developing highly accurate models for monkeypox diagnosis, such as "PoxNet22," which achieved 100% precision, recall, and accuracy [13]. Additionally, emotion classification from social media posts using deep learning models like CNN, Long-Short Term Memory (LSTM) and Bi-Directional LSTM (BiLSTM) have provided insights into public sentiment and concerns regarding monkeypox, indirectly aiding in geographical classification by identifying areas of heightened concern [13]. Deep-learning methods supported with transfer learning tools and hyperparameter optimization have been employed to detect monkeypox through skin lesions, with models like MobileNetV3-s showing remarkable results [14]. The development of an image-based deep convolutional neural network, MPXV-CNN, for identifying characteristic skin lesions caused by monkeypox, has shown robust classification performance across various skin tones and body regions [15]. Ensemble learning-based frameworks that combine probabilities from pre-trained base learners like Inception V3, Xception, and DenseNet169 have also been proposed to detect monkeypox virus presence from skin lesion images with high accuracy [16]. Mobile applications using deep learning for preliminary diagnosis of monkeypox through skin lesion images offer a quick and accessible tool for individuals, potentially aiding in the geographical classification by facilitating early detection [17]. Longitudinal studies assessing spatiotemporal risk factors of monkeypox infection and predicting global epidemiological trends using modified SEIR models and k-means clustering analysis have provided insights into changing risk factors and future outbreak predictions [18]. Finally, the comparison of different pre-trained deep learning models fine-tuned for monkeypox virus detection has led to the development of ensemble approaches that improve overall performance, aiding health practitioners in mass screening [19]. These methods collectively contribute to the geographical classification of monkeypox disease by enabling accurate and early detection, understanding public sentiment, and predicting future outbreaks, thereby assisting in containment efforts and public health planning.

The fact that deep learning is successful in various fields and is effective in the classification of monkeypox disease geographically, led us to analyze the DNA data of monkeypox disease with KANs, a new deep learning model, in this study. The utilization of color-coded representations in genetic sequence visualization offers a compelling approach to elucidate complex genomic information. By assigning specific colors to each nucleotide—blue for Adenine, yellow for Thymine, red for Cytosine, green for Guanine, and white for undefined nucleotides—researchers can create an intuitive visual map of DNA structures, facilitating the identification of patterns and motifs within the genetic code. This method is further enhanced by the implementation of advanced edge detection algorithms, which employ brightness and geometric thresholds, along with segmentation area parameters, to accentuate structural features that might otherwise remain obscure in traditional digital mapping techniques. Moreover, these color-coded images are designed to serve as input for transfer learning classification models, effectively bridging the gap between bioinformatics and machine learning. This innovative approach enables the application of sophisticated pattern recognition and classification techniques to genetic sequence data through their visual representations, potentially offering new insights into genomic structures and functions.

In the study, DNA sequences associated with monkeypox disease were initially obtained from both American and Asian regions. These sequences were subsequently color-coded, and DNA motifs were generated. Following motif creation, a filtering process was conducted, and the resultant images were subjected to classification using both KANs and various Artificial Neural Network (ANN) models, which were then compared. There are several deficiencies of traditional spatial transcriptomics (ST) methods such as

the use of tools like SPADE and PERSIST, and the application of graph-based ML methods, are associated with the paper's approach. SPADE and PERSIST tools exemplify tailored approaches developed to leverage the unique aspects of ST data. The paper's use of KANs can be seen as a continuation of this trend, aiming to provide superior performance in detecting spatially relevant gene expression patterns and selecting informative gene targets. Also in Graph-Based ML methods, The application of graph-based ML methods in previous research highlights the potential of combining ST with ML to uncover spatially restricted niches and signaling networks. This paper's approach with KANs aligns with this by offering a more nuanced understanding of disease pathology at the spatial level, potentially contributing to new therapeutic strategies.

Model performance was evaluated based on precision, recall, and F1-score metrics. The highlights of this study can be explained as follows:

- For the first time in this study, the DNA sequences that cause monkeypox disease were color-coded.
- To the best of our knowledge, for the first time in this study, KANs deep learning model was employed to classify monkeypox DNA motifs.
- This study shows that deep learning-based image-screen methods can be effective in regional transcriptome and phylogenetic analysis.

The remainder of the study is organized as follows: In the second section, information about the data set and methods used in the study is given. In the third section, application results are given, and the findings obtained with both KANs, and other models are compared. In the fourth section, discussion is made and the performance of the KANs model is examined. In the last section, the study was concluded, suggestions and future studies were mentioned.

Material and Methods

Datasets

The DNA sequences utilized in the research were sourced from GISAID [20] and the National Center for Biotechnology Center (NCBI Virus) [21]. These databases offer a comprehensive analysis of Human Monkey Pox DNA records globally since its isolation in December 2022. During the investigation, we acquired the Asian-tagged series from the NCBI Virus database and the remaining sequences from GISAID. Data archived and shared in these repositories are formatted as FASTA [22] files, containing essential details such as date, location, quality, and publication information of the researchers involved in isolating the virus's DNA sequence. The repositories also indicate the quality of DNA sequences, leading us to avoid utilizing incomplete or low-quality sequences in our classification model. Presently, GISAID has published over 5,000 complete or partial genomic sequences, while the NCBI Virus database contains more than 2,000 fully or partially labeled sequences. The cumulative count of complete gene sequences obtained from these databases amounts to 3,165. The distribution of these sequences is as follows: Europe 900, America (North and South) 1,448, Asia 926, as depicted in Figure 1. The data integrity and completeness in these repositories are signified by the requirement of a base pair count exceeding 29,000 and an unresolved amino acid percentage of less than 5%. These criteria ensure that the DNA information collected accurately represents all amino acid values. Despite our preference for fully isolated gene sequences, occasional utilization of sequences containing "N" placeholders was necessary due to limited data availability. Due to the relative unevenness in the distribution of DNA sequences, the data set collected during the study was divided into two groups: American and Asian. In such a study, a simple binary classification problem was applied. This resulted in two almost equally distributed classes.

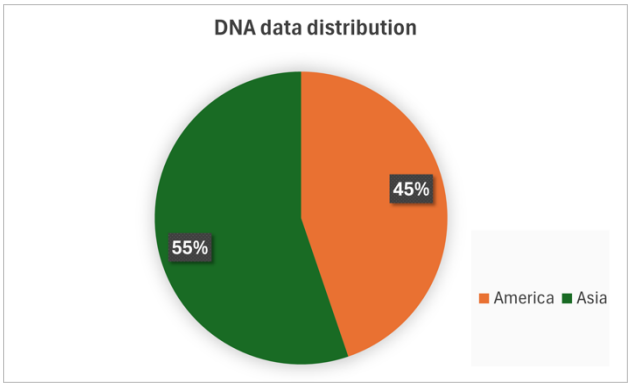

Figure 1. The complete genome sequences counts are distributed by continents for the training dataset.

DNA Motifs

When the DNA sequences of the Human Monkey Pox virus were transformed into visual representations, a strategic approach was employed by drawing an analogy and portraying them in the form of circular shapes. It is worth noting that an DNA sequence is comprised of four fundamental nucleobases, namely Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). The representation of these nucleobases is further enhanced by the utilization of distinct color codes; Adenine-Yellow, Cytosine-Blue, Guanine-Green and Thymine-Red.

The algorithm employed for the creation of DNA motifs is a sophisticated computational procedure utilized for the precise determination of the multitude of points necessary to discretize a circular shape into pixels. Within the intricate flow of this algorithmic process, the initial step involves the identification of points within the eight divided segments, followed by the meticulous determination of points within the remaining octant. In the meticulous process of pinpointing each point (x, y) along the circumference of the circle, the subsequent pixel coordinates are calculated as either (x, y + 1) or (x-1, y + 1), ensuring a systematic approach to pixel placement.

The application of this algorithm was pivotal in the task of populating the circle generated through the utilization of

DNA data, meticulously translated into a character array. This method resulted in the creation of intricate motifs, whereby the RNA sequences were transformed and represented in images with a resolution of 200x200 pixels and 3 color channels, thereby enhancing the visual representation of the genetic information. The varying lengths of gene sequences under scrutiny necessitated a meticulous determination of the optimal dimensions within our circle drawing algorithm, ensuring a standardized and efficient computational process. The flowchart of the process of the stages here is given in Appendix-A in detail. The detailed process steps of the flow chart are as follows:

The algorithm for DNA motif creation and filtering represents a sophisticated approach to transforming genetic sequence data into visual representations suitable for advanced analytical techniques, including machine learning applications. This process encompasses multiple stages, from initial data preprocessing to the generation of filtered images, each step carefully designed to ensure the integrity and relevance of the resulting motifs.

The process commences with the initialization phase, wherein the algorithm establishes crucial file paths for FASTA input, genome storage, motif output, and filtered motif storage. Concurrently, a regex filter is prepared to facilitate efficient text processing. Subsequently, the algorithm engages in a meticulous FASTA file processing stage. Here, the contents of the FASTA file are read and parsed using regex splitting, effectively separating the file into discrete entries for individual analysis.

Following the initial parsing, the algorithm proceeds with a line-by-line examination of the genetic sequences. This stage implements a series of stringent filters to ensure only high-quality, relevant genetic data is processed further. The filtering criteria include checks for non-empty lines, exclusion of header lines (those starting with ">"), minimum length requirements (greater than 50 characters), and the absence of long stretches of undefined nucleotides ("NNNNNNNN"). Lines meeting these criteria contribute to the construction of a comprehensive genome string, which is periodically written to storage and checked against a minimum length threshold to ensure sufficient genetic material for meaningful analysis.

The core of the motif creation process lies in the image generation phase. Here, the algorithm initializes key parameters such as image radius and origin, creating a bitmap with dimensions of 200x200 pixels. The algorithm then employs a mathematical approach to determine the height of each column in the circular representation, calculated as the square root of the difference between the squared radius and the squared x-coordinate. This calculation ensures a proper circular shape in the resulting image.

In the pixel processing stage, the algorithm iterates through each calculated y-coordinate, mapping nucleotides from the genetic sequence to specific colors: Adenine to blue, Thymine to yellow, Cytosine to red, Guanine to green, and any undefined nucleotides to white. This color-coding scheme creates a visually distinct representation of the genetic sequence, with each pixel in the image corresponding to a specific nucleotide in the original sequence.

Post-generation, the algorithm applies a sophisticated edge detection technique to enhance the visual patterns within the motif. This process involves the careful setting of brightness and geometric thresholds, as well as a segmentation area parameter. The edge detection algorithm examines each pixel in the context of its local neighborhood, defined by a circular mask. By comparing brightness values and calculating a segmentation area, the algorithm determines whether each pixel represents an edge or a continuous region, thereby highlighting the structural features of the genetic sequence in the visual representation.

The final stages of the algorithm involve the storage of both the original motif image and its edge-detected variant, saved as PNG files in their respective directories. These images serve as the end product of the visualization process, encapsulating complex genetic information in a format conducive to further computational analysis.

Ultimately, these generated images are primed for utilization in transfer learning classification models. This final step bridges the gap between bioinformatics and machine learning, allowing for the application of advanced pattern recognition and classification techniques to genetic sequence data through their visual representations.

In conclusion, this algorithm represents a multifaceted approach to genetic sequence visualization, combining elements of bioinformatics, image processing, and machine learning preparation. By transforming complex genetic data into standardized, visually interpretable formats, it paves the way for novel insights and analytical approaches in genomic research and related fields.

Throughout the process of motif creation, any gaps or unoccupied pixels were elegantly filled with white color to seamlessly complete the circular shape, albeit with a subtle linear flaw discernible on the right periphery. From the perspective of artificial neural networks, these images can be interpreted as multi-dimensional arrays with dimensions of 200x200x3, encapsulating the intricate details of the DNA motifs. The resultant DNA motif, a product of this intricate algorithmic process, is visually presented in Figure 2 showcasing the culmination of computational precision and biological data integration. Before obtaining the images in this particular context, it was necessary for us to utilize the FASTA files that were acquired from various datasets, a crucial step in the process aimed at transforming these files into motifs. This transformation involved the meticulous separation of the data contained within these files into distinct DNA sequences, a task that was accomplished through the implementation of a specialized application that we meticulously crafted using the Delphi programming language.

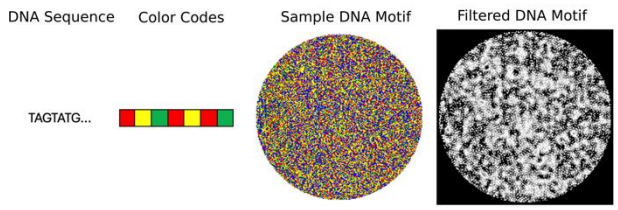

Figure 2. The motif of the sample isolated Human Monkey Pox DNA and generated and filtered nucleobase the motif.

Based on the results of this preliminary examination, clusters were created using the K-means clustering algorithm in combination with the Principal Component Analysis method. The aim here is to determine whether it is possible to separate the motif images in a phylogenetic analysis. The application showed that a total of 49 clusters belonging to 3165 DNA motifs were formed. Cluster 5 and cluster 21 were selected and analyzed as examples from the clusters obtained here. The first results obtained can be seen in Figure 3.

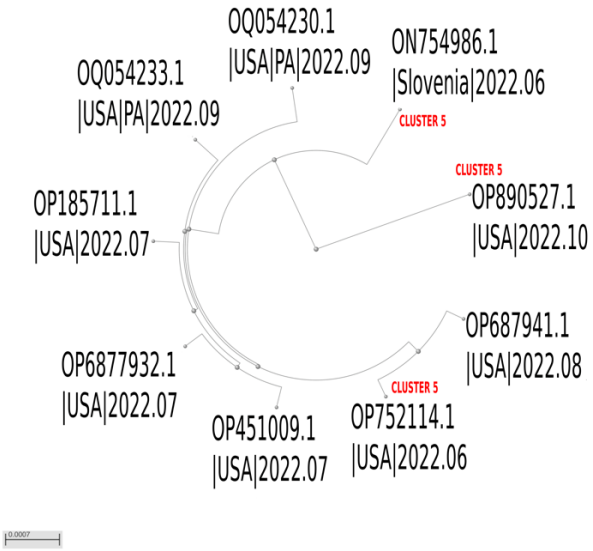

Figure 3. K-means sample cluster comparing with phylogenetic distributions.

The genomes shown in Figure 3 with accession numbers OP890527, OP752114, ON754986 belong to cluster 5 and the others to cluster 21. This random selection suggests that the DNA-motif application can be used to classify genomes to a certain extent.

Also when observing the motif files acquired through our research endeavors, it became evident that certain segments within them exhibit a repetitive nature. Upon closer examination, we assessed that these reiterated sections potentially signify specific patterns within the DNA configuration of the virus, whereby some segments remain unaltered while others undergo mutations. The accurate functioning of the classification model hinges on the ability to recognize and delineate these patterns, particularly crucial for pinpointing mutations that vary across different geographical regions. Furthermore, we deliberated that the identification of these patterns could prove advantageous in constructing the hierarchical structure of phylogenetic trees using newly obtained DNA sequences, as well as in drawing comparisons with previously analyzed genetic sequences.

A prominent technique employed in the realm of visual information processing, including tasks like image segmentation and pattern recognition, pertains to the utilization of the edge detection methodology. Over time, numerous algorithms have been devised for edge detection purposes. Nevertheless, it is noteworthy that the inception of low-level applications marked the initial stages of image processing methodologies, which have progressively evolved to more sophisticated levels in contemporary times. In our quest to identify repetitive motifs within the motif files at hand, we opted for the utilization of the low-level edge detection algorithm. Applying the image filter converted RGB pictures to grayscale from motifs obtained. Picture dimensions are stored as 200x200x3.

Kolmogorov-Arnold Networks

Kolmogorov-Arnold Networks (KANs) emerge as highly promising alternatives to Multi-Layer Perceptrons (MLPs) within the realm of neural networks. It is important to note that KANs boast robust mathematical underpinnings akin to those of MLPs: the latter are established upon the foundational universal approximation theorem, whereas the former find their basis in the esteemed Kolmogorov-Arnold representation theorem [24]. In a fascinating duality, KANs and MLPs exhibit contrasting characteristics: KANs implement activation functions on edges, whereas MLPs employ activation functions on nodes. This seemingly subtle alteration actually renders KANs superior to MLPs in terms of both model accuracy and interpretability [25].

KANs, highlights the foundational role of MLPs in deep learning, acknowledged for their expressive power in approximating nonlinear functions, as guaranteed by the universal approximation theorem. However, MLPs for their significant drawbacks, including their consumption of a vast majority of non-embedding parameters in models like transformers and their relative lack of interpretability without the aid of post-analysis tools. In contrast, KANs are proposed with learnable activation functions on edges, replacing every weight parameter with a univariate function parametrized as a spline, which leads to improvements in accuracy and interpretability over MLPs. KANs, with their architecture, mathematical foundation, and potential for scientific discovery, promising a significant leap in accuracy and interpretability for data fitting and PDE solving while potentially overcoming the curse of dimensionality.

KANs leveraging the Kolmogorov-Arnold representation theorem to propose a neural network architecture with learnable activation functions on edges, replacing traditional weight parameters with univariate functions parametrized as splines. The methodology begins with the design of a neural network that explicitly parametrizes the Kolmogorov-Arnold representation, using B-spline curves with learnable coefficients for each 1D function, forming the basis of KANs. This approach allows for the creation of a prototype KAN, visualized as a two-layer neural network with activation functions placed on edges and simple summation performed on nodes. To enhance the model's capability, the paper discusses generalizing KANs to be wider and deeper, addressing the challenge of extending the Kolmogorov-Arnold representation to deeper networks by drawing an analogy between MLPs and KANs and defining a "KAN layer" as a matrix of 1D functions with trainable parameters. Furthermore, KANs with existing methods, highlighting the continuous learning and robustness of KANs over traditional symbolic regression techniques. It also delves into the scaling laws and intrinsic dimensionality, providing a theoretical framework for understanding the efficiency and effectiveness of KANs in terms of model parameters and test loss. The main differences between KAN networks and standard MLP networks are shown in Table 1.

Table 1. Kolmogorov-Arnold Networks (KANs) vs. Multi-Layer Perceptrons (MLPs).

Kolmogorov-Arnold Network (KAN)	Multi-Layer Perceptron (MLP)
$f(x) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \Phi_{q.p}(x_p) \right)$	$f(x) \approx \sum_{i=1}^{N(\varepsilon)} a_i \sigma(w.x + b_i)$
Sum operation on nodes and learnable activation functions on edges.	Learnable weights on edges and fixed activation functions on nodes.
$\mathrm{KAN(x)} = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(x)$	$\mathrm{MLP(x)} = (w_3 \circ \sigma_2 \circ w_2 \circ \sigma_1 \circ w_1)(\mathrm{x})$

In Table 1, univariate functions $\Phi_{p,q}$, Φ_q defined as $\Phi_{p,q} : [0, 1] \rightarrow R$ and $\Phi_q : R \rightarrow R$. Given an input vector x_0, in a network of L KAN layers, KAN(x) is the output of the network.

Experiments and Results

Before the classification with KAN networks, existing methods were tested in order to correctly classify the motif structure designed for this study. One of the problems that can be encountered in DNA classification studies is that the random evolution process in gene changes is not predictable. However, after the motifs are obtained, the first naive examination shows that some patterns follow each other. This view is that they can be analyzed by basic classification methods. For this purpose, the previously proven transfer learning method was applied. Transfer learning plays a pivotal role in the field of artificial intelligence (AI) by enabling the application of knowledge gained from one task to improve performance on a related but different task. This technique is particularly beneficial in scenarios where labeled data for the target task is scarce, allowing models to leverage larger datasets from related tasks to overcome overfitting and enhance performance on the target task. distinguishes itself from traditional machine learning techniques through its unique approach of leveraging pre-existing knowledge from one task to improve performance on a related, yet distinct, task [29-31]. Unlike conventional machine learning methods that start the learning process from scratch for each new task, transfer learning capitalizes on the insight gained from previously solved problems to enhance learning efficiency and accuracy for new problems. This is particularly advantageous in scenarios where labeled data for the new task is scarce or expensive to obtain, as it allows the model to bypass the intensive data requirement typically necessary for training machine learning models from the ground up. Moreover, transfer learning is versatile in its application, encompassing a range of computational intelligence-based techniques, including neural networks, evolutionary algorithms, swarm intelligence, and fuzzy logic, to improve performance further than what vanilla transfer learning can achieve on its own [29].

Transfer learning operates by leveraging the knowledge acquired from one or more source tasks to improve the learning efficiency and performance on a related target task. This process is particularly beneficial in scenarios where labeled data for the target task are scarce or expensive to obtain. At its core, transfer learning involves two main stages: pre-training and fine-tuning [30]. During the pre-training stage, a model is trained on a source task that has abundant labeled data. This model learns a set of features or representations that are potentially useful for the target task. For instance, in the domain of circuit performance prediction, neural networks optimally trained on data from one technology node can learn features that are transferable to another technology node, significantly reducing the amount of data required for accurate predictions in the target node [31].

In this study, the weight values extracted from the pre-existing networks were transferred initially to an AveragePooling layer. Subsequently, a 50% dropout was implemented to transfer the values from this layer to the neural network connections, thus mitigating overfitting during the learning process. The dropout mechanism facilitates the removal of certain network cells from the model, consequently averting overfitting of the neural network. During the final phase, the DNA sequences were directed to the 4-dimensional fully connected layer designated for classifying the four classes. The non-linear function selected for this layer was SoftMax. SoftMax function operates by taking a vector of K real numbers as input and normalizing it to a distribution of K probabilities proportionate to the exponents of the input numbers. Additionally, within the model, the loss function employed was categorical cross-entropy. This function is commonly utilized for single label categorization, signifying that only one class is relevant for each data point. Optimization method was preferred and used as RMSProp (Root Mean Square Propagation) for training. Optimization functions are used to determine the learning rate of the artificial neural network. The Learning Rate value in the optimization function was set as 0.001. Also training lasted for 15 epochs.

The classification results of the pre-trained networks used in the study are shown in Table 2. The network structure chosen for the best classification reflects an optimal model.

Table 2. Test dataset results obtained in various artificial neural networks.

#	Model	Precision	Recall	f1-score	Test Accuracy
0	MobileNet	0.6855	0.6571	0.62056	0.6571
1	MobileNetV2	0.6468	0.6444	0.6247	0.6444
2	InceptionV3	0.6383	0.6222	0.5768	0.6222
3	ResNet50	0.7054	0.5809	0.4506	0.5809
4	ResNet101	0.3157	0.5619	0.4043	0.5619
5	DenseNet121	0.5265	0.5619	0.4043	0.5619
6	VGG16	0.3157	0.5619	0.4043	0.5619
7	InceptionResNetV2	0.3157	0.5619	0.4043	0.5619
8	VGG19	0.3157	0.5619	0.4043	0.5619
9	DenseNet169	0.6302	0.6190	0.5764	0.6190

Table 2 shows that the MobileNet network gives the best results. Here, the pruned weights of this network are thought to cause a more effective classification. The comparative analysis of various artificial neural networks reveals MobileNet as the optimal model, demonstrating superior performance across multiple evaluation metrics. MobileNet achieved the highest recall and test accuracy (both 0.6571), indicating its proficiency in correctly identifying positive instances and overall classification accuracy. Its precision (0.6855) was second only to ResNet50, showcasing its ability to minimize false positives. The F1-score (0.62056), a harmonic mean of precision and recall, further corroborates MobileNet's balanced performance. These metrics are derived from standard formulae:

Precision = TP / (TP + FP),

Recall = TP / (TP + FN),

F1 = 2 * (Precision * Recall) / (Precision + Recall),

Accuracy = (TP + TN) / (TP + TN + FP + FN),

where TP, FP, TN, and FN represent True Positives, False Positives, True Negatives, and False Negatives, respectively. MobileNet's consistent high performance across these metrics underscores its efficacy in scenarios requiring a balance between precision and recall, coupled with high overall accuracy. After this stage, DNA sequences that were separated from the dataset and not included in the training were tested with MobileNet. In this network, Test Loss is 0.5982, Test accuracy is 65.71% and Cohen Kappa Score is 0.25737. Other results of the network are as shown in Table 3.

Table 3. MobileNet results of Test dataset.

Region	Precision	Recall	f1-Score	Support
America	0.75	0.33	0.45	138
Asia	0.64	0.92	0.75	177

Accuracy and loss graphs for training set obtained is shown in Figure 4.

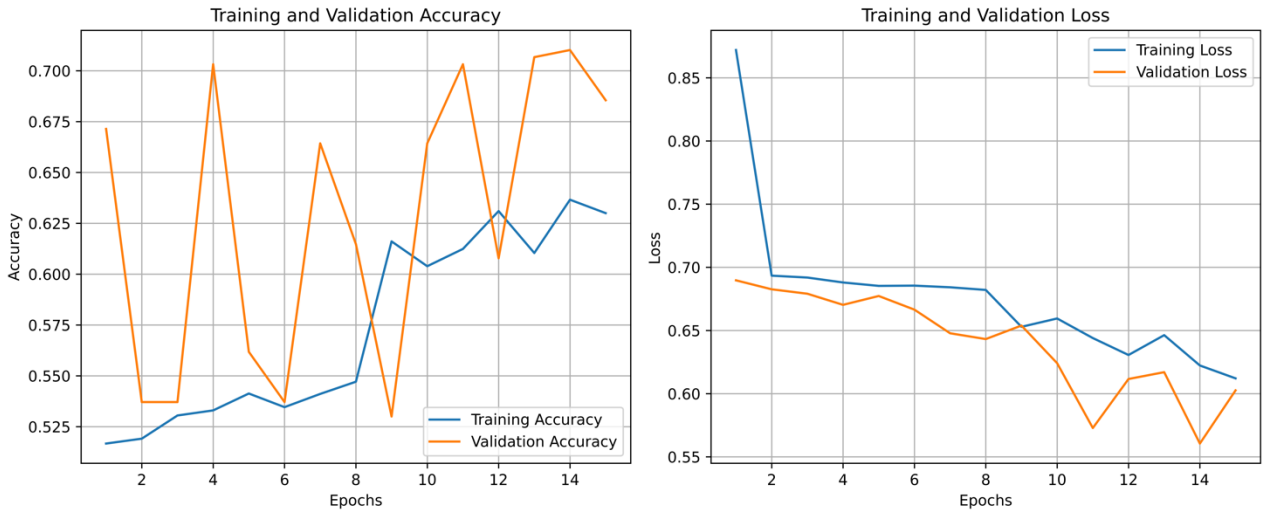

Figure 4. Training loss and accuracy plots of MobileNet model.

The confusion matrix of the test results is also shown in Figure 5. The confusion matrix reveals a notable level of accuracy in classifying our motif files. Despite the minor inaccuracies, it is essential to acknowledge the interrelation among the data elements, such as the mutation-induced connection observed in virus DNA. The contentious nature of this assertion deems it as a focal point for further discussion. Extensive investigation is necessary to support the claim that the error matrix also highlights these distinctions, particularly due to the absence of absolute geographical delineations in virus mutations.

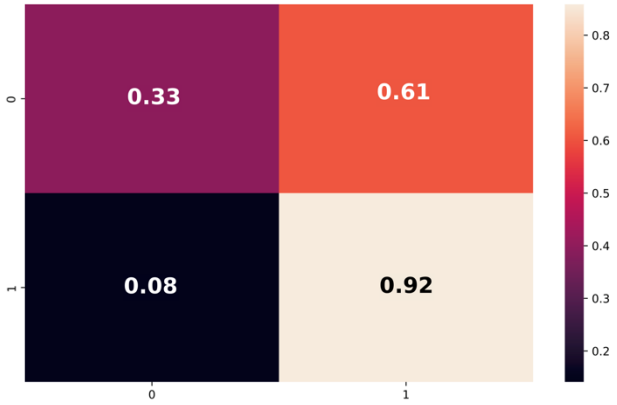

Figure 5. Confusion matrix was obtained in the test of two classes (America:1, Asia 0) in MobileNet model.

Considering the success of the network model used in this study in the transfer learning process, the equivalent structure was applied to KAN networks. However, as expected, the results were not obtained as expected since the KAN network did not have a set of weights whose success was calculated in advance. Accuracy and loss plots for training process shown in Figure 6.

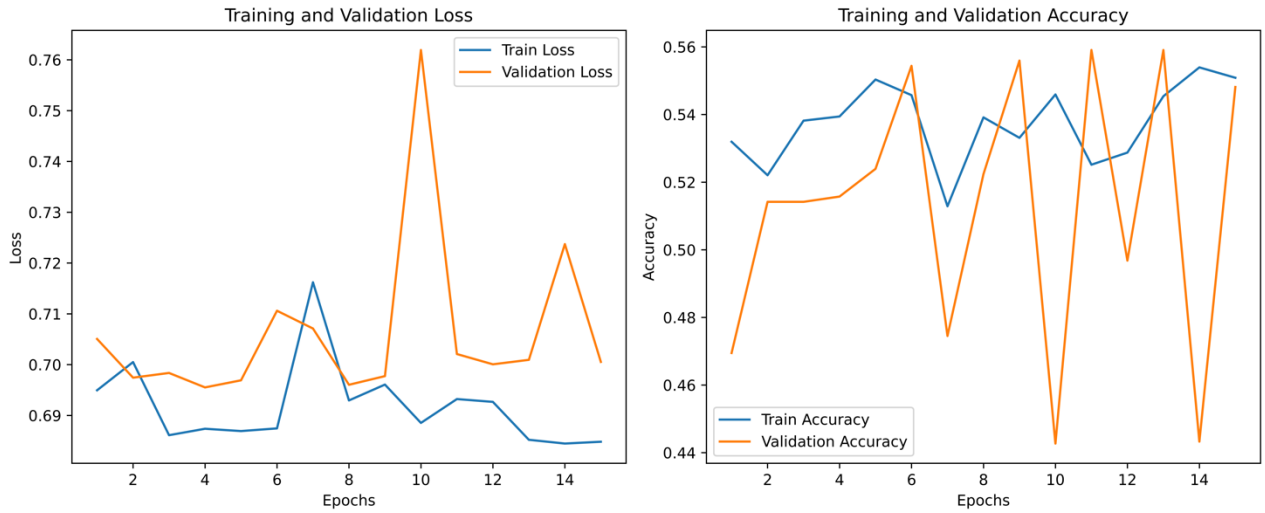

Figure 6. Training loss and accuracy plots of KAN model

In KAN networks, sum functions are used from the weights. In various experiments, an effective sum function could not be found. However, since it is still a new technology, the results obtained are still promising. According to the results obtained from the test data Accuracy is 0.45182, Precision is 0.43129, Recall is 0.45182, F1 Score is 0.32187 and Cohen Kappa Score is -0.02551. In Table 4, observed values from the results obtained on the test data set are shown.

Table 4. KAN results of Test dataset.

Region	Precision	Recall	f1-Score	Support
America	0.45	0.93	0.61	292
Asia	0.41	0.04	0.07	341

The confusion matrix of the test results in KAN is also shown in Figure 7.

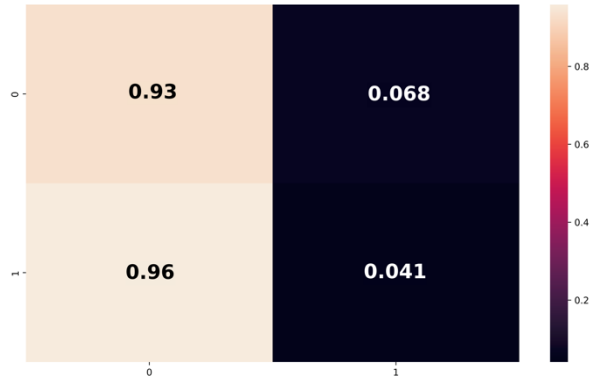

Figure 7. Confusion matrix was obtained in the test of two classes in KAN (America:1, Asia 0).

Discussion and Conclusion

DNA classification is a critical component in the fields of bioinformatics and computational biology, contributing significantly to genome annotation, disease diagnosis, and evolutionary studies. Through the process of genome annotation, DNA classification facilitates the identification of functional elements within genomes, such as genes, regulatory regions, and non-coding DNAs. This is essential for comprehending the organization and function of genomes, thereby providing foundational insights for further genetic research. In the realm of disease diagnosis and prognosis, accurate DNA classification aids in the detection of genetic variations associated with various diseases. This capability enables early diagnosis and prognosis, thereby allowing for the development of personalized treatment strategies that can improve patient outcomes. Spatial transcriptomics analysis is a powerful technique for understanding tissue heterogeneity and gene expression patterns within their spatial contexts. By preserving the spatial organization of cells within a tissue, spatial transcriptomics allows researchers to identify distinct cell types, their spatial arrangement, and their interactions. This is crucial for deciphering the complexity of tissue composition and function. Furthermore, spatial transcriptomics has the potential to uncover novel cell types and states along with their spatial relationships, thus providing a more detailed understanding of tissue architecture and dynamics. In the study of disease pathology, spatial transcriptomics is instrumental in examining gene expression changes within the context of disease progression. Despite the significant advancements, there are several challenges associated with employing deep learning techniques in DNA classification and spatial transcriptomics analysis. One major challenge is the limited availability of labeled data, which is essential for training deep learning models. The acquisition of labeled DNA sequences or spatial transcriptomics data is often expensive, time-consuming, and necessitates expert annotation. Another challenge is class imbalance within genomic and transcriptomic datasets, where some classes, such as rare genetic variants or cell types, are underrepresented. This imbalance can lead to biased models that do not perform well on minority classes. Additionally, the length and complexity of DNA sequences pose difficulties for deep

learning models, which must capture long-range dependencies and intricate patterns to be effective. As the field advances, deep learning is anticipated to play an increasingly vital role in unraveling the complexities of genomes and spatial gene expression patterns, ultimately contributing to significant progress in basic research, disease understanding, and personalized medicine.

According to the KAN results obtained in the study, for the America region, the model achieves a Precision of 0.45, indicating that 45% of the instances predicted as America are actually from America. The Recall is 0.93, which means that the model correctly identifies 93% of all instances that truly belong to America. The F1-Score of 0.61 suggests a moderate balance between Precision and Recall. The Support value of 292 represents the total number of instances from America in the dataset. On the other hand, the model's performance for the Asia region is significantly lower. The Precision is 0.41, meaning that 41% of the instances predicted as Asia are correctly classified. However, the Recall is only 0.04, indicating that the model identifies just 4% of all instances that actually belong to Asia. This extremely low Recall suggests that the model is struggling to recognize instances from Asia, leading to a high number of false negatives. The F1-Score of 0.07 further confirms the poor performance for this region. The Support value of 341 shows that there are more instances from Asia than America in the dataset.

The observed performance gap between the two regions implies a potential bias towards the America region within the model, as indicated by the notably higher Recall metric. Several factors could contribute to this bias:

First, an imbalanced dataset might be a contributing factor, wherein the training data comprises a disproportionately larger number of instances from the America region compared to Asia. This imbalance can lead the model to favor the majority class, potentially skewing its performance towards better recognition of instances from the overrepresented region. Second, discrepancies in feature representation could exacerbate the bias. It is plausible that the features utilized for classification exhibit greater discriminative power for instances originating from the America region, thus facilitating easier identification by the model. This could stem from inherent differences in the characteristics or distributions of features between the two regions. Lastly, the selected model architecture might inherently predispose towards capturing patterns specific to the America region more effectively. Certain architectural choices, such as network depth, layer configurations, or activation functions, could inadvertently favor learning representations that align better with the characteristics prevalent in the America region, consequently amplifying the observed bias in model performance.

Addressing these potential sources of bias necessitates careful consideration during the model development and evaluation process. Strategies for mitigating bias include ensuring balanced representation of instances from different regions in the training data, augmenting features to enhance their discriminative power across diverse regions and exploring alternative model architectures that are more agnostic to regional disparities. By adopting such approaches, the model's robustness and generalizability across varied geographical contexts can be enhanced, thereby fostering more equitable performance outcomes.

Despite the impressive results achieved so far, there is still significant room for improvement and further development of KANs. One area of active research is the exploration of new network architectures that can better capture the intricate relationships within data. This includes the investigation of deeper and more complex network structures, as well as the incorporation of attention mechanisms and memory components. By designing more sophisticated architectures, researchers aim to unlock the full potential of KANs and push the boundaries of their performance. Another promising avenue for future development is the integration of KANs with other machine learning techniques. For instance, combining KANs with deep learning approaches, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), could lead to powerful hybrid models that leverage the strengths of both paradigms. Additionally, the incorporation of transfer learning and multi-task learning strategies could enable KANs to efficiently learn from related tasks and domains, further enhancing their adaptability and generalization capabilities. Furthermore, the interpretability and explainability of KANs are crucial aspects that require further investigation. While these networks have shown remarkable performance, understanding the internal representations and decision-making processes of KANs remains a challenge. Developing techniques to visualize and interpret the learned features and decision boundaries of KANs will not only improve their trustworthiness but also facilitate their application in domains where transparency is essential, such as healthcare and finance.

In conclusion, Kolmogorov-Arnold Networks have emerged as a promising state-of-the-art technology in the field of machine learning. The current results obtained using KANs are highly encouraging, showcasing their ability to learn and represent complex functions efficiently. However, there is still significant potential for further development and improvement. By exploring new network architectures, integrating KANs with other machine learning techniques, and addressing interpretability and explainability challenges, researchers can unlock the full potential of these networks. As the field of machine learning continues to evolve, it is expected that KANs will play an increasingly important role in shaping the future of artificial intelligence and its applications across various domains.

Acknowledgement

The GISAID and NCBI(Virus) researchers did not participate in the analysis or writing of this paper.

References

[1] H. Park et al., “Spatial Transcriptomics: Technical Aspects of Recent Developments and Their Applications in

Neuroscience and Cancer Research," Adv. Sci., vol. 10, no. 16, p. 2206939, Jun. 2023, doi: 10.1002/advs.202206939.

[2] A. A. Heydari and S. S. Sindi, "Deep learning in spatial transcriptomics: Learning from the next next-generation sequencing," Biophys. Rev., vol. 4, no. 1, p. 011306, Mar. 2023, doi: 10.1063/5.0091135.

[3] D. F. Miyagishima et al., "157 Identifying Spatial Transcriptomics Signaling Networks in Human Glioblastoma Using Graph-Based Machine Learning," Neurosurgery, vol. 69, no. Supplement_1, pp. 42–42, Apr. 2023, doi: 10.1227/neu.0000000000002375_157.

[4] I. Covert, R. Gala, T. Wang, K. Svoboda, U. Sümbül, and S.-I. Lee, "Predictive and robust gene selection for spatial transcriptomics," Nat. Commun., vol. 14, no. 1, p. 2091, Apr. 2023, doi: 10.1038/s41467-023-37392-1.

[5] A. J. Lee, R. Cahill, and R. Abbasi-Asl, "Machine Learning for Uncovering Biological Insights in Spatial Transcriptomics Data," 2023, doi: 10.48550/ARXIV.2303.16725.

[6] Z. Qiu, S. Li, M. Luo, S. Zhu, Z. Wang, and Y. Jiang, "Detection of differentially expressed genes in spatial transcriptomics data by spatial analysis of spatial transcriptomics: A novel method based on spatial statistics," Front. Neurosci., vol. 16, p. 1086168, Nov. 2022, doi: 10.3389/fnins.2022.1086168.

[7] F. Qin, X. Luo, B. Cai, F. Xiao, and G. Cai, "Spatial pattern and differential expression analysis with spatial transcriptomic data," Jul. 09, 2023. doi: 10.1101/2023.07.06.547967.

[8] A. Robles-Remacho*, R. M. Sanchez-Martin, and J. J. Diaz-Mochon*, "Spatial Transcriptomics: Emerging Technologies in Tissue Gene Expression Profiling," Apr. 28, 2023. doi: 10.26434/chemrxiv-2023-n20f0.

[9] M. Zahmatyar et al., "Human monkeypox: history, presentations, transmission, epidemiology, diagnosis, treatment, and prevention," Front. Med., vol. 10, p. 1157670, Jul. 2023, doi: 10.3389/fmed.2023.1157670.

[10] Y. Li, S. Stanojevic, and L. X. Garmire, "Emerging artificial intelligence applications in Spatial Transcriptomics analysis," Comput. Struct. Biotechnol. J., vol. 20, pp. 2895–2908, 2022, doi: 10.1016/j.csbj.2022.05.056.

[11] M. M. Ahsan et al., "Deep transfer learning approaches for Monkeypox disease diagnosis," Expert Syst. Appl., vol. 216, p. 119483, Apr. 2023, doi: 10.1016/j.eswa.2022.119483.

[12] O. Attallah, "MonDiaL-CAD: Monkeypox diagnosis via selected hybrid CNNs unified with feature selection and ensemble learning," Digit. Health, vol. 9, p. 205520762311800, Jan. 2023, doi: 10.1177/20552076231180054.

[13] F. Yasmin et al., "PoxNet22: A Fine-Tuned Model for the Classification of Monkeypox Disease Using Transfer Learning," IEEE Access, vol. 11, pp. 24053–24076, 2023, doi: 10.1109/ACCESS.2023.3253868.

[14] R. Olusegun, T. Oladunni, H. Audu, Y. Houkpati, and S. Bengesi, "Text Mining and Emotion Classification on Monkeypox Twitter Dataset: A Deep Learning-Natural Language Processing (NLP) Approach," IEEE Access, vol. 11, pp. 49882–49894, 2023, doi: 10.1109/ACCESS.2023.3277868.

[15] M. Altun, H. Gürüler, O. Özkaraca, F. Khan, J. Khan, and Y. Lee, "Monkeypox Detection Using CNN with Transfer Learning," Sensors, vol. 23, no. 4, p. 1783, Feb. 2023, doi: 10.3390/s23041783.

[16] A. H. Thieme et al., "A deep-learning algorithm to classify skin lesions from mpox virus infection," Nat. Med., vol. 29, no. 3, pp. 738–747, Mar. 2023, doi: 10.1038/s41591-023-02225-7.

[17] R. Pramanik, B. Banerjee, G. Efimenko, D. Kaplun, and R. Sarkar, "Monkeypox detection from skin lesion images using an amalgamation of CNN models aided with Beta function-based normalization scheme," PLOS ONE, vol. 18, no. 4, p. e0281815, Apr. 2023, doi: 10.1371/journal.pone.0281815.

[18] V. H. Sahin, I. Oztel, and G. Yolcu Oztel, "Human Monkeypox Classification from Skin Lesion Images with Deep Pre-trained Network using Mobile Application," J. Med. Syst., vol. 46, no. 11, p. 79, Oct. 2022, doi: 10.1007/s10916-022-01863-7.

[19] J. Gao et al., "Monkeypox outbreaks in the context of the COVID-19 pandemic: Network and clustering analyses of global risks and modified SEIR prediction of epidemic trends," Front. Public Health, vol. 11, p. 1052946, Jan. 2023, doi: 10.3389/fpubh.2023.1052946.

[20] Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data – from vision to reality," Eurosurveillance, vol. 22, no. 13, Mar. 2017, doi: 10.2807/1560-7917.ES.2017.22.13.30494.

[21] E. L. Hatcher et al., "Virus Variation Resource – improved response to emergent viral outbreaks," Nucleic Acids Res., vol. 45, no. D1, pp. D482–D490, Jan. 2017, doi: 10.1093/nar/gkw1065.

[22] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison.," Proc. Natl. Acad. Sci., vol. 85, no. 8, pp. 2444–2448, Apr. 1988, doi: 10.1073/pnas.85.8.2444.

[23] W. S. Klug, M. R. Cummings, C. A. Spencer, M. A. Palladino, and D. J. Killian, Essentials of genetics, Tenth edition. Hoboken, NJ: Pearson, 2020.

[24] A. N. Kolmogorov, "On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition," presented at the Doklady Akademii Nauk, Russian Academy of Sciences, 1957, pp. 953–956.

[25] Z. Liu et al., "KAN: Kolmogorov-Arnold Networks," ArXiv Prepr. ArXiv240419756, 2024.

[26] E. Waisberg et al., "Transfer learning as an AI-based solution to address limited datasets in space medicine," Life Sci. Space Res., vol. 36, pp. 36–38, Feb. 2023, doi: 10.1016/j.lssr.2022.12.002.

[27] L. Jin, C. Qu, Y. Zhang, C. Fan, Z. Zhu, and S. Liu, "Transfer Learning on Trial: A Case Study to Apply Existing Models to Heterogeneous Datasets," in 2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms (PRMVIA), Beihai, China: IEEE, Mar. 2023, pp. 292–296. doi: 10.1109/PRMVIA58252.2023.00054.

[28] K. Combs, H. Lu, and T. J. Bihl, "Transfer Learning and Analogical Inference: A Critical Comparison of Algorithms, Methods, and Applications," Algorithms, vol. 16, no. 3, p. 146, Mar. 2023, doi: 10.3390/a16030146.

[29] J. Wang and Y. Chen, "From Machine Learning to Transfer Learning," in Introduction to Transfer Learning, in Machine Learning: Foundations, Methodologies, and Applications. , Singapore: Springer Nature Singapore, 2023, pp. 39–52. doi: 10.1007/978-981-19-7584-4_2.

[30] A. H. Ali, M. G. Yaseen, M. Aljanabi, S. A. Abed, and C. Gpt, "Transfer Learning: A New Promising Techniques," Mesopotamian J. Big Data, pp. 29–30, Feb. 2023, doi: 10.58496/MJBD/2023/004.

[31] Z. Wu and I. Savidis, "Transfer Learning for Reuse of Analog Circuit Sizing Models Across Technology Nodes," in 2022 IEEE International Symposium on Circuits and Systems (ISCAS), Austin, TX, USA: IEEE, May 2022, pp. 1033–1037. doi: 10.1109/ISCAS48785.2022.9937457.

APPENDIX-A: Flowchart of motif generation and classification process