

The process of DNA classification involves organizing genetic material into taxonomic groups based on shared characteristics and evolutionary relationships. Chaos Game Representation (CGR) has been successfully utilized in conjunction with various Machine Learning algorithms to achieve high accuracies in DNA sequence classification. However, working with CGR-based feature vectors can be computationally demanding, leading to challenges in efficient classification due to the resulting high-dimensional feature space. Our goal is to identify a subset of CGR-based features through applying various dimensionality reduction techniques on the feature space while maintaining similar classification performance and enhancing computational efficiency. To assess the effectiveness of our approach, we employed a novel alignment-free supervised machine learning pipeline to classify over 15,500 complete mitochondrial genomes sourced from the NCBI reference dataset, spanning from the Kingdom to the Genus levels. Initially, we computed k-mer frequencies for the genomic sequences using CGRs, followed by the derivation of corresponding magnitude spectra by applying the Discrete Fourier Transform to the flattened CGRs. Subsequently, Machine learning algorithms were trained on the feature vectors generated by performing Principal Component Analysis (PCA) to the magnitude spectra. Results demonstrated that a subset of features identified through PCA achieved classification performance that was not only comparable but, in certain instances, even superior to that attained with the complete feature space. Furthermore, these selected features exhibited improved computational efficiency. Interestingly, we observed that even a few components, enough to explain 25% of the total variance, yielded accuracy levels equivalent to or surpassing those achieved with the entire feature space. In summary, this research highlights the potential of employing PCA for dimensionality reduction while preserving crucial features which enhances the efficiency and accuracy of the classification. Additionally, our approach provided valuable insights into the genomic datasets by identifying the most significant features for classification.