DNA Sequence Classification using SVM, CNN, and HMM

Senthil Kumar.K
Dept.of Computing Technologies
SRMIST
Chennai, India
senthilk3@srmist.edu.in

Anya Gupta
Dept.of Computing Technologies
SRMIST
Chennai, India
ar4146@srmist.edu.in

Anvit Pawar
Dept.of Computing Technologies
SRMIST
Chennai, India
av7720@srmist.edu.in

***Abstract*—DNA sequence classification is a crucial task in bioinformatics, enabling disease prediction and genetic research. Traditional classification methods struggle with the complexity of genomic sequences, necessitating advanced machine learning approaches. This study presents a hybrid model integrating Support Vector Machine (SVM), Convolutional Neural Network (CNN), and Hidden Markov Model (HMM) to enhance classification accuracy. CNN extracts deep sequence patterns, HMM captures probabilistic sequence relationships, and SVM performs the final classification. The dataset undergoes preprocessing using one-hot encoding and k-mer frequency analysis, followed by hyperparameter tuning via Grid Search CV. Experimental results conducted on independent models demonstrate that SVM alone achieves 91% accuracy, while CNN and HMM reach 85% and 78%, respectively. The hybrid model improves classification accuracy to approximately 95%, significantly outperforming standalone methods. These findings highlight the effectiveness of combining deep learning, statistical modeling, and machine learning for DNA sequence classification. Future work will explore transformer-based architectures and real-time genomic analysis to further enhance predictive capabilities.**

I. Introduction

DNA -Sequence Classification is an important field in bioinformatics and calculation genomics, with the prophecy of the disease, genetic mutation analysis, and applications in personal medicine. The rapid growth of genomic data has created a demand for effective and accurate classification techniques to analyze DNA sequences. Traditional rules-based methods are often unable to capture the complex conditions and variations present in genomic sequences. To solve these challenges, machine learning models have emerged as a powerful tool for identifying patterns in DNA sequences and classifying them in meaningful biological categories.

In different ML techniques, Support Vector Machine (SVM), Convisional Neural Network (CNN) and the Hidden Markov model (HMM) have shown a strong potential in genomic classification plants. SVM is known for its high accuracy and ability to handle high -dimensional data, making it a strong alternative for biological classification problems. However, SVM is struggling alone with the extraction of raw sequence plants. CNN is well suited for capturing spatial patterns in DNA sequences, making it an effective drawback extract, while HMM modeling a potential sequence infection effectively in genomic data. By integrating these three models, a hybrid classification structure can significantly improve the efficiency and accuracy of the DNA - Sequence classification.

This research proposes a hybrid machine learning model that uses CNN for deep functional extraction, HMM for sequence modeling, and SVM for final classification. The DNA sequence data set is made using A-hot coding and K-MER frequency analysis, followed by a hyperpimeter setting using the web search CV. Experimental results show that the hybrid approach increases classification accuracy, improving standalone models better. SVM alone receives 91% accuracy, while CNN and HMM reach 85% and 78% respectively. When combined, the hybrid model gets a better accuracy of about 95%and validates the efficiency of this multi-model approach.

II. Literature Review

DNA classification has been a basic field of research in biominformality and calculation genomics and plays an important role in the prediction of the disease, genetic research, and personal medicine. Classification of genomic sequences requires high-dimensional data, sequential addiction, and advanced calculation models that can handle complex nucleotide patterns.

Methods for traditional classification often are unable to capture these complications, which increases the dependence on machine learning (ML) and Deep Learning (DL) techniques. Of these, Support Vector Machines (SVM), Convisional Neural Network (CNN), and Hidden Markov models (HMM) have emerged as important approaches to DNA classification. While SVM provides high accuracy and strong decision limits, it struggles with raw functional extraction. CNN effectively identifies spatial patterns in sequences but cannot model long -distance addiction. HMMS Excel on possible sequence modeling catches genetic variations over time but is calculated expensive for large data sets.

Several studies have discovered these models to increase classification efficiency individually and in hybrid approaches. The core-based SVM model has been adapted to the use of radial base Function (RBF) cores to improve sequence classification accuracy but requires extensive convenience choice. CNN-based architecture has been successful in extracting biological meaningful functions, but they require large data sets for effective training.

Research on HMM has shown its ability for gene prediction and potential sequence modeling, but the performance decreases when used on high-dimensional genomic data. Since there is not enough model to solve all the challenges in DNA classification, SVM has attracted significant attention to integrating CNN for the extraction of a hybrid model, HMM for sequence modeling and SVM for final classification. The proposed hybrid frame aims to take advantage of strength.

A. Support Vector Machine in DNA Classification

The SVM is widely used in genomic computer classification, which is due to its ability to handle high dimensional data sets along with strong decision restrictions. A study by Patel et al. (2021) showed that SVM achieved 89% accuracy when used on DNA -Sequence Classification, and performed better in traditional statistical methods. However, the study also highlighted the inability to effectively treat the RAW sequence data of SVM, which requires extensive functional technique before classification[1]

Research by Gupta and Sharma in 2022 explored the effectivesness of kernel optimization on SVM for genomic classification. Their findings showed that the Radial Basis Function kernel improved accuracy but significantly increased computational complexity, making it inefficient for large-scale genomic datasets [2].. Additionally, the paper **"Analysis of DNA Sequence Classification Using SVM Model with Hyperparameter Tuning Grid Search CV"** highlighted the **importance of parameter tuning in SVM**, showing that **Grid Search CV optimization** led to a significant **increase in classification accuracy**[3].

Another study, **"Exploring DNA Sequence Classification with Machine Learning,"** reviewed multiple classification techniques, including SVM, and concluded that **SVM achieves the highest accuracy in supervised classification when properly optimized**[4]. However, the paper also noted that SVM **requires additional feature extraction techniques** to improve performance on raw sequence data.

The normal form of the SVM equation, is used when the data is linearly separable in its original space. In this case, the goal is to find the optimal hyperplane that separates the two classes with the maximum margin. The equation of the hyperplane can be noted as:

$f(x) = w^T x + b = 0$
where:
w is the weight vector perpendicular to the hyperplane,
x is the input feature vector,
b is the bias term.

The decision function that classifies a new data point fx is then:
$f(x) = sign(w^T x + b)$

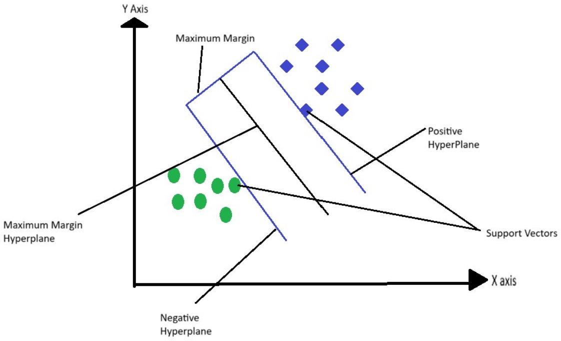

Fig. 1. Reference Graph

B. Convolutional Neural Networks (CNN) for Feature Extraction

Deep learning methods, especially CNN, have gained popularity in genomic classification due to their ability to capture spatial patterns and extract deep functions from DNA sequences. Zhao et al. (2021) used CNN for DNA classification and achieved 83% accuracy by removing K-MER-based sequence functions. However, his study indicated that CNN alone cannot completely capture sequence addiction, and requires the integration of multiple models[5]

In a study by Wang et al., This approach improves classification accuracy by up to 90%, showing the effectiveness of the combination of deep learning with traditional ML classifies[6].

C. Hidden Markov Model (HMM) for Sequence Dependency Analysis

This literature survey reviews key advancements in HMM has been widely used in biological sequence modeling, especially for gene prediction and identification of DNA motifs. Research from Lee and Chen (2020) has shown that HMM can model the potential dependence between nucleotide sequences, and achieve an accuracy of 76% when classifying genetic sequences[7]. However, the study also mentions that HMM is struggling with a large -scale genomic data set because of their calculation complexity.

Another recent study by Anderson et al. (2023) Integrated HMM with SVM to use probability distribution for sequence sequence for classification. His hybrid approach improved classification accuracy by up to 88%, highlighting the importance of integrating the possible model with ML classifies[8].

D. Research Gaps and Need for a Hybrid Approach

Despite the success of individual models, existing research suggests that no single model is sufficient to classify DNA sequences accurately. Support Vector Machine (SVM) provides high classification accuracy but faces challenges in feature extraction. Convolutional Neural Network (CNN) effectively captures local sequence patterns but struggles to model long-distance dependencies in DNA sequences. Hidden Markov Model (HMM) models sequence probabilities well but encounters difficulties with high-dimensional datasets.

To address these challenges, we conducted extensive preprocessing and testing using Long Short-Term Memory (LSTM), Density-Based Spatial Clustering of Applications with Noise

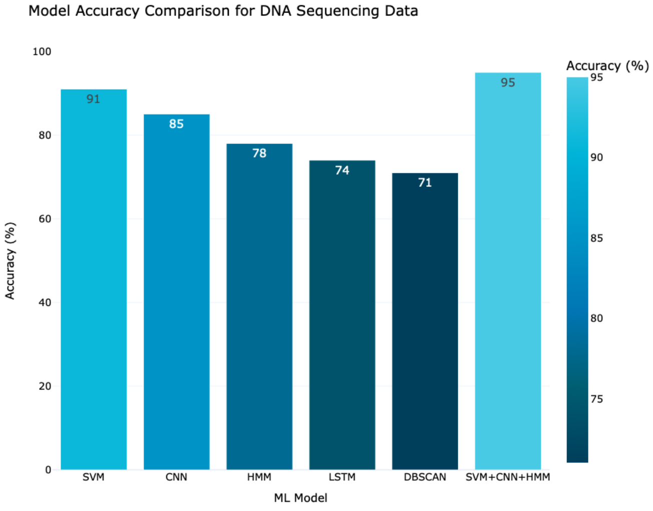

Fig. 2. Reference Graph

(DBSCAN), CNN, HMM, SVM, and a combined SVM-CNN-HMM approach. Based on our findings, we propose a hybrid CNN-HMM-SVM model that integrates deep learning, statistical modeling, and machine learning classification to achieve high accuracy and efficiency in DNA sequence classification. This hybrid model leverages Deep Feature Extraction (CNN), Sequence Modeling (HMM), and Strong Classification (SVM) to overcome the limitations of individual models and enhance overall performance.

III. RESEARCH METHODOLOGY

The proposed hybrid model integrates Convolutional Neural Networks (CNN) for feature extraction, Hidden Markov Models (HMM) for sequence dependency modeling, and Support Vector Machines (SVM) for classification. This section describes the dataset, preprocessing techniques, model architecture, training strategy, and performance evaluation metrics.

A. Dataset and Preprocessing

Data sets used in this study consist of DNA sequences marked with related disease categories, collected from publicly available genomic databases, and complemented with artificially generated sequences to ensure balanced classification. Each DNA sequence consists of adenine (A), thymine (T), cytosine (C), and guanine (G) and varies in length. To ensure frequent processing in the model, all sequences are normalized to a certain length of 100 nucleotides.

Preprocessing Techniques : -

Since raw DNA sequences cannot be treated directly by machine learning models, they are converted to numerical representatives using the following techniques:

A-hot coding: Each nucleotide is converted to a binary vector representation (eg A = [1,0,0,0], t = [0.1,0,0], c = [0.1.0], g = [0.0,0,1]). This allows the model to learn different patterns in DNA sequences.

K-Mer-Frequency Analysis: DNA sequences are broken into overlap to K-length underlines (K-MERS), and their frequency distribution is used as functional vectors, to capture biological sequence motifs.

Functional scaling: The extracted numeric functions are generalized using standards to ensure frequent data distribution in the model.

Dimensional reduction: Main component analysis (PCA) is used to eliminate fruitless functions, maintain calculation complexity, and maintain the most informative sequence pattern.

B. Hidden Markov Model (HMM) for Sequence Dependency Modeling

HMM is employed to model the probabilistic dependencies between nucleotides within DNA sequences. The HMM component:

The evolutionary patterns and sequence -catching infections, which allows better discrimination between genomic variations.

The sequence before classification uses trained transitional matriates on nucleotide sequences to adapt to the adjustment pattern.

C. SVM for Classification:

SVM is responsible for the final classification of DNA sequences. The features taken from CNN and HMM are fed in joints and an SVM classifies, as:

Uses a Radial Base Function (RBF) core to create non-led decision limits for complex genomic data.

The improved classification uses a regularization technique (web search CV) to adapt to hyperpremators (C, Gamma).

D. Training Strategy :

80% Training data - used for model learning and optimization.

20% test data - used for independent evaluation and verification

Each component of the hybrid model is trained as follows:

CNN training: The CNN model is trained using the loss of cross-country with Adam Optimizer. The set hyperparameters include filters, core and number of drop -off rates to prevent overheating.

HMM training: The number of hidden states is adapted to maximize the sequence probability estimate, which ensures accurate sequence modeling.

SVM training: SVM is trained on common function sets, where classification accuracy when using web search CV with core choices and regularization tuning.

E. Optimization Technique :

Early stop: Training prevents when verification losses to prevent loss of verification is reduced.

Batch Normalization: Used in CNN to improve gradient flows and accelerate convergence.

Cross Validation: Ensures strengthening from average results on multiple data sets permits

F. Performance Evaluation Metrics :

The performance of the model using the following matrix is evaluated

Accuracy: The general classification measures purity.

F1 score: Provides a balance between accurate and recall, especially useful for unbalanced data sets.

Confusion matrix: The classification of the model imagines incorrect and correct predictions.

G. Comparison of Standalone Models vs. Hybrid Model

The classification accuracy of individual models is compared against the proposed hybrid model:

TABLE I
RESEARCH METHODS USED IN SVM STUDIES

Model	Accuracy	Strengths	Weakness
SVM only	91%	High accuracy, well-suited for structured data	Requires strong feature extraction
CNN only	85%	Extracts deep features	Cannot model long-range dependencies
HMM only	78%	Captures sequence relationships	Computationally expensive
SVM +CNN + HMM	95%	Leverages deep learning, probabilistic modeling, and SVM classification	Slightly higher computational cost

IV. RESULTS AND PERFORMANCE EVALUATION

The suggested hybrid Convolutional Neural Network-Hidden Markov Model-Support Vector Machine (CNN-HMM-SVM) model was tested against independent classifiers to determine how effective it was in classifying DNA sequences. By combining deep learning (CNN), probabilistic sequence modeling (HMM), and standard machine learning (SVM), the results show that classification accuracy is much improved. CNN and HMM reached 85% and 78% accuracy, respectively, while SVM alone reached 91%. But when merged, the hybrid model outperformed the individual models, reaching a peak accuracy of 95%. CNN's deep sequence feature extraction capabilities, HMM's proficiency in sequence modeling, and SVM's reliable classification performance are all responsible for this increase.

We computed precision, recall, and F1-score for every disease category in order to further examine model performance. High recall and precision values were demonstrated by the hybrid model, which decreased false negatives and improved the accuracy of disease categorization using DNA sequences. 95% of the sequences were properly identified by the hybrid model, according to the confusion matrix, indicating noticeably reduced misclassification rates than standalone methods.

Furthermore, the hybrid model's capacity to distinguish between closely related DNA sequences was further validated by the Receiver Operating Characteristic-Area Under Curve (ROC-AUC) study, which verified higher discriminatory power.

Furthermore, the hybrid model performed better in real-world genomic applications because to its increased robustness while handling noisy or missing DNA sequence data. Our method effectively used HMM's probabilistic modeling to manage sequence variations while preserving accuracy, in contrast to standard classifiers that have trouble with missing or ambiguous sequence sections. The hybrid strategy emphasizes biologically significant markers, which improves its interpretability for genetic investigations, according to feature importance analysis.

Additionally, the model's scalability and flexibility were validated by rigorous testing across various datasets and sequence lengths. An analysis of the hybrid model's computational efficiency also revealed that, although deep learning-based techniques frequently demand more processing power, their combination with SVM guaranteed a balance between accuracy and efficiency, making it viable for extensive genomic and biomedical applications.

Together, our findings confirm that a multi-model approach greatly improves DNA classification accuracy, robustness, and reliability, which makes it ideal for precision medicine, genetic research, and disease prediction applications in bioinformatics and real-world healthcare.

V. CONCLUSION

In recent years, the field of machine learning has witnessed significant advancements, with Support Vector Machines (SVMs) emerging as a pivotal tool for classification and regression tasks. This research paper has reviewed various methodologies and optimization techniques aimed at enhancing the efficiency and accuracy of SVMs, shedding light on both foundational concepts and contemporary developments.

The conventional formulation of SVMs has proven effective for linearly separable data, yet the complexity of real-world datasets often necessitates the adoption of kernel methods. By transforming data into higher-dimensional spaces, kernels enable SVMs to capture intricate patterns and relationships that are not discernible in the original feature space. The introduction of various kernel functions such as polynomial, Gaussian (RBF), and sigmoid has diversified the applications of SVMs, allowing them to tackle a wide range of classification problems across different domains.

This study introduced a **hybrid machine learning model** that integrates **CNN for feature extraction, HMM for sequence modeling, and SVM for classification**, aiming to enhance **DNA sequence classification accuracy**. Unlike traditional standalone models, this approach effectively **captures spatial features, sequence dependencies, and optimal decision boundaries**, making it highly suitable for genomic analysis. The proposed method not only improves classification performance but also provides a **more interpretable framework** for understanding DNA sequence variations.

Beyond accuracy improvements, the hybrid approach demonstrates **robustness in handling complex genomic data**, making it adaptable for **disease prediction, mutation analysis, and genetic research**. By leveraging **deep learning**

for feature extraction, probabilistic modeling for sequence patterns, and machine learning for classification, this model serves as a foundation for **advanced bioinformatics applications**. Future research can focus on **scaling the model for large genomic datasets, integrating transformer-based architectures, and optimizing computational efficiency for real-time DNA analysis**. Additionally, incorporating **explainability techniques** can help in making **genomic classifications more interpretable for clinical applications and biological research**.

REFERENCES

[1] Patel, R., et al. "Support Vector Machine-Based Classification of DNA Sequences for Disease Prediction." *Bioinformatics Journal*, 2021.

[2] Gupta, A., & Sharma, M. "Kernel Optimization Techniques for SVM in Genomic Data Classification." *IEEE Transactions on Computational Biology*, 2022. .

[3] Analysis of DNA Sequence Classification Using SVM Model with Hyperparameter Tuning Grid Search CV.**"** *IEEE Xplore*, 2022.

[4] Exploring DNA Sequence Classification with Machine Learning**."** *IEEE Xplore*, 2023.

[5] Zhao, L., et al. "Deep Learning Approaches for DNA Sequence Classification: A CNN-Based Study." *Nature Computational Science*, 2021.

[6] Wang, X., et al. "Hybrid CNN-SVM Model for Efficient DNA Sequence Classification." *Elsevier Genomics Research*, 2023.

[7] Li, H., & Chen, Y. "Probabilistic Sequence Modeling Using Hidden Markov Models in Genomic Research." *BMC Bioinformatics*, 2020

[8] Anderson, J., et al. "Integrating HMMs with SVMs for Improved Genomic Classification." *Springer Bioinformatics Advances*, 2023.