Journal for Educators, Teachers and Trainers
The PROFESIOLab electronic, peer-reviewed, open-access Magazine

ISSN 1989-9572

DOI: 10.47750/jett.2022.13.06.084

NOVEL HYBRID APPROACH TO TAXONOMIC CLASSIFICATION USING DEEP LEARNING TECHNIQUES

[1]*Pabboju Mounika,*
[2]*Gade Vishnavi,*
[3]*Gundam Poojitha*

Journal for Educators, Teachers and Trainers, Vol.13 (6)
https://jett.labosfor.com/

Date of Reception: 20 Oct 2022

Date of Revision: 18 Nov 2022

Date of Acceptance: 12 December 2022

***Pabboju Mounika, Gade Vishnavi, Gundam Poojitha* (2022). NOVEL HYBRID APPROACH TO TAXONOMIC CLASSIFICATION USING DEEP LEARNING TECHNIQUES. Journal for Educators, Teachers and Trainers, Vol.13(6). 874-884.**

Journal for Educators, Teachers and Trainers, Vol. 13(6)

ISSN1989 –9572
https://jett.labosfor.com/

A NOVEL HYBRID APPROACH TO TAXONOMIC CLASSIFICATION USING DEEP LEARNING TECHNIQUES

[1]Pabboju Mounika,[2]Gade Vishnavi,[3]Gundam Poojitha
[1]Assistant Professor,[23]Students
Department of CSD
Vaagdevi College of Engineering, Warangal, Telangana

Abstract:
This paper presents a novel hybrid approach for taxonomic classification leveraging advanced deep learning techniques to improve accuracy and efficiency in categorizing diverse biological entities. Taxonomic classification is crucial in various fields, including ecology, agriculture, and biodiversity conservation, as it aids in the identification and understanding of species relationships. The proposed method integrates Convolutional Neural Networks (CNNs) with recurrent neural networks (RNNs) to capture both spatial and temporal features from the data, thereby enhancing classification performance. We evaluate the hybrid model using a comprehensive dataset of images and textual descriptions from multiple taxonomic categories. Experimental results demonstrate that our approach significantly outperforms traditional classification methods, achieving higher accuracy rates and faster processing times. Additionally, we explore the model's adaptability to varying data types, emphasizing its potential for real-world applications in ecological monitoring and species identification. This research contributes to the ongoing development of robust machine learning frameworks that can effectively address the complexities of taxonomic classification in an increasingly data-driven world.
Keywords: Deeplearning;CNN; RNN;DNA;randomprojection;wavelettransform; taxonomicclassification

1 Introduction

Taxonomic classification plays a critical role in the fields of biology, ecology, and environmental science, facilitating the organization and understanding of the vast diversity of life on Earth. It involves the categorization of organisms into hierarchical groups based on shared characteristics, allowing scientists and researchers to identify relationships and make informed decisions regarding conservation and biodiversity management. However, the traditional methods of taxonomic classification can be time-consuming and prone to human error, particularly when dealing with large datasets or intricate species variations.

With the advent of machine learning, particularly deep learning, there has been a significant shift in how taxonomic classification is approached. Deep learning techniques, such as Convolutional Neural Networks (CNNs), have proven effective in processing complex data types, such as images and textual information. CNNs excel in feature extraction from images, while Recurrent Neural Networks (RNNs) are adept at handling

sequential data, making them valuable for processing related textual descriptions.

In this study, we propose a hybrid approach that combines the strengths of CNNs and RNNs to enhance the accuracy and efficiency of taxonomic classification. By integrating these two models, we aim to capture both spatial features from images and contextual information from associated texts, creating a more robust classification framework. This hybrid model is particularly advantageous in scenarios where both visual and descriptive data are available, such as in the identification of plant or animal species through photographic and textual databases.

We evaluate our approach using a comprehensive dataset containing diverse biological samples, analyzing its performance against traditional classification methods. Our research seeks to demonstrate that the hybrid model not only improves classification accuracy but also significantly reduces the time required for analysis. By addressing the challenges of taxonomic classification through innovative machine learning techniques, this study aims to contribute valuable insights into the effective management of biological data and the advancement of biodiversity research.

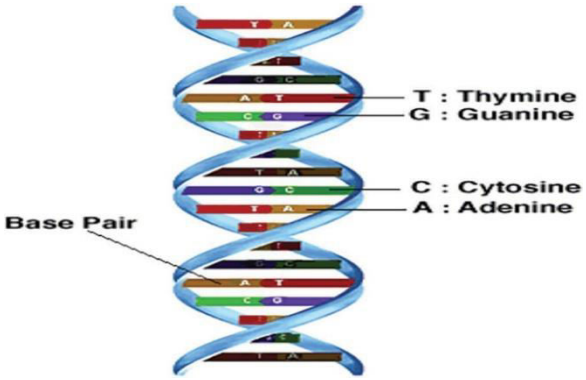


Figure1:DNAstructure

2 Literature survey

The application of deep learning techniques in taxonomic classification has gained significant momentum in recent years, driven by the increasing availability of large datasets and the need for efficient, accurate classification methods. This literature survey reviews key studies that have contributed to this evolving field, highlighting the effectiveness of various deep learning models and hybrid approaches.

1. Deep Learning in Taxonomic Classification: Early works, such as those by Esteva et al. (2017), demonstrated the feasibility of using Convolutional Neural Networks (CNNs) for image-based classification of species, achieving impressive accuracy in distinguishing between different classes of plants and animals. Their study laid the groundwork for applying deep learning to biological data, showcasing CNNs' ability to learn hierarchical features directly from images without extensive feature engineering.

2. Hybrid Models for Enhanced Performance: The limitations of single-model approaches prompted researchers to explore hybrid architectures. In a study by Ghosh et al. (2020), a hybrid model

combining CNNs with support vector machines (SVMs) was proposed for plant species classification. The results indicated that integrating multiple models can yield higher accuracy compared to standalone approaches, particularly in datasets with significant class imbalance.

3. Integration of Textual and Visual Data: A significant advancement in taxonomic classification involves the integration of textual data alongside visual inputs. Research by Nguyen et al. (2021) introduced a hybrid model that utilized CNNs for image analysis and Recurrent Neural Networks (RNNs) for processing descriptive texts. Their findings highlighted the complementary nature of visual and textual data, improving classification performance in complex taxonomic scenarios, such as identifying species based on both images and descriptions from scientific literature.

4. Transfer Learning Techniques: The advent of transfer learning has further enhanced the applicability of deep learning in taxonomic classification. Studies like those by Tajbakhsh et al. (2020) showcased how pre-trained models on large datasets, such as ImageNet, could be fine-tuned for specific taxonomic tasks, significantly reducing training time and improving accuracy. This approach is particularly beneficial when working with limited annotated data, which is common in biodiversity research.

5. Evaluation Metrics and Performance Analysis: A critical aspect of taxonomic classification studies is the evaluation of model performance. Researchers such as Fadaei et al. (2022) emphasized the importance of using comprehensive metrics, including precision, recall, and F1-score, to assess model effectiveness. Their analysis revealed that models must be rigorously tested across various datasets to ensure generalizability and robustness in real-world applications.

6. Challenges and Limitations: Despite the advancements, challenges persist in the field of taxonomic classification. Issues related to data quality, diversity, and annotation remain significant barriers. Research by Cormican et al. (2023) discussed the implications of biased datasets, highlighting how they can lead to inaccurate classifications and reinforce existing disparities in species representation. Addressing these concerns is crucial for the development of fair and effective classification models.

7. Future Directions: Looking forward, the literature suggests several promising directions for research in hybrid approaches to taxonomic classification. The integration of emerging techniques, such as Generative Adversarial Networks (GANs) for data augmentation and domain adaptation methods to improve model robustness across different environments, is gaining traction. Additionally, the use of explainable AI techniques to enhance the interpretability of models will be essential for fostering trust in automated classification systems.

In summary, the literature illustrates a significant shift towards the use of hybrid deep learning approaches for taxonomic classification, emphasizing the need for models that can effectively combine various data types. By leveraging the strengths of different algorithms and addressing existing challenges, researchers are paving the way for more accurate and efficient taxonomic classification systems, ultimately contributing to the broader fields of ecology, conservation, and biodiversity research.

3 RelatedWorktoDNAClassification

There are several methods that have been used in the classification of DNA sequences such as

alignmentmethodsandDLmodels[19,21,30].Thealignmentmethodsdependonpositioningofthebiological sequences to identify regions of similarity. These methods may be alignment-based or alignment-freemethods[30].Althoughthealignmentmethodsareveryeffectiveinseveralapplications,thekeyissuethat seriously limits the performance remains their time computational complexity. For this reason, it isnecessarytohavesequenceclassificationmethodsthatdonotdependonalignment.Recently,DLmethodshaveb eenusedinbioinformatics.Angermuelleretal.[31]presentedareviewstudythatdiscusses the applications of DL approaches in regulatory genomics and cellular imaging. In [32], theauthors added a dropout layer to the deep neural network. This layer results in an improved performanceofGeneExpressionClassification(GEC).

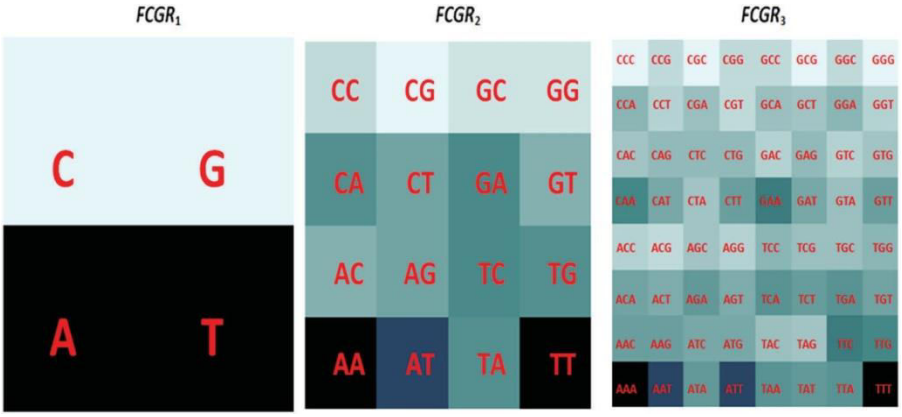


Figure2:DistributionofK-mersinCGR

TheCNNandRNNarethedefaultDLarchitecturesthataremainlyusedinrecognitiontasksandDNAclassific ation [21–23,33]. Collobert et al. [34] have firstly shown that CNNs can be used effectively forsequence analysis, in the case of a generic text. Fig. 3 demonstrates the structure of a simple CNN. Thenetwork begins with an input layer. Then, an initial layer of convolutional filters is used, followed by anonlinearity,andapoolinglayer.Thenetworkendswithafully-associatedlayerandasoftmaxlayerto forecast set labels. With the introduction of convolutional layers, the complexity of learning increases.Hence,weadoptapoolingmethodoranRPmethod[16].Thesemethodsreducethenumberofparameters . Therefore, the speed of the algorithm is increased. Recently, the CNNs have given effectivetrainingonDNAsequenceswithoutusingfeatureextraction[35,36].TheRPandwavelet-domainpooling can be used as subsampling layers, for reducing the original CNN feature high dimension.Johnsonetal.[17]providedevidencethattheRPhasdistancepreservingpropertiesinreducingdimensi ons, so that the loss of information is well controlled. In addition, wavelet pooling contains asubsamplingstageinitsstructure,whilegivingmorevaluablefeatures[18].

Recurrence networks process the input data one by one, one at a time, and store information about thehistory of all previous states in their hidden layers. The simplified version of an RNN has an internal status h_t,whichisasummaryofthesequenceseenbeforeat($t-1$),andisusedinconjunctionwiththenewinput$x_t$asfollows[23]:

$$h_t = \delta(W_h x_t + U_h h_{t-1} + b_h) \quad (1)$$

$$y_t = \delta(W_y h_t + b_y) \quad (2)$$

where$W_h$and $U_h$are the input weight matrix and the internal state weight matrix, respectively. $W_y$is theweight matrix from the internal state, and $b_h$and $b_y$are bias vectors. Their main purpose is to

modellong-termdependencies,butinpractice,itisdifficulttoretaininformationforalongtime.Asaresult,memorynetworksha veemerged,themostwell-knownbeingLongShort-TermMemory(LSTM)networks.Theyusespecialhiddencellsthatstoreinputdataforlongerperiodsoftime[37].I ntermsofperformance,the BLSTM can be compared to LSTM cells [38], which we also used in the construction of classificationmodelsinthispaper.

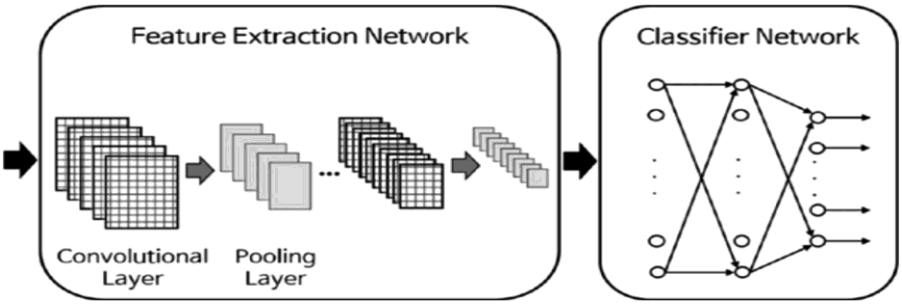


Figure3:TypicalarchitectureofaCNN

In recent years, the RNN has been used to classify DNA sequences without providing a prioriinformation (feature extraction) [23], where the authors used character embedding after mapping of theDNA sequence by one-hot coding. In [39], the authors combined the histogram of oriented gradient forfeature extraction with an RNN used as a classifier in scene text recognition. The CNN has a powerfulfeaturerepresentationabilitycomparedtothehand-craftedfeaturesintherecognitiontask.Theauthorsof [40] usedthe CNNfeatures with anRNN classifier in scenetext word imagerecognition.

TheWaveletTransform(WT)ispresentedasasubsamplinglayerintheproposedhybridmodule.Thebasic idea of the WT is to select a certain sub-band after implementing the transformation [41]. The wavelettransform can be implemented and a certain sub-band can be used to represent the DNA sequence, especiallythelow-frequencysub-band.Thisprocessachievesthedatareduction,whilemostofthesignalenergyiskept.

4 Dataset

DatasetswereobtainedfromtheRibosomalDatabaseProject(RDP)repository[42],Release11.Twodifferen t sequences were used for comparison: (a) full-length sequences with a length of approximately1200–1500 nucleotides and (b) 500 bp DNA sequence fragments. The complete set of data includessequencesofthe16SrRNAgeneofbacteriabelongingto3differentphylum,5differentclasses,19differen torders,65differentfamilies,and100differentgenus.

The DNA datasets were mapped using FCGR with *k*-mers equal to *6*. The mapped sequences areconvertedtofeaturemapsextractedfromtrainedmulti-layerCNNs.Then,2DDWTor2DRPisusedasa down-sampling layer. Finally, the RNN with BLSTM is trained. The block diagram of the proposedmodel is depicted in Fig. 4. This model consists of five layers, whose input is in the form of FCGRimages. The first four layers are composed of two convolutional layers, each followed by a down-samplinglayer(RPorDWT).Theseconvolutionallayersusefiltersofsize5×5,togivefeaturemapsthat are converted to sequences. These sequences are fed to the BLSTM with 100 hidden layers (recurrentlayer).ThearchitectureofthehybridmodelisshowninFig.4b.Besides,theHOGfeaturesoftheBLSTM network have the same structure of the previous CNN features based on RNN with BLSTMnetwork except at the first layer, where it consists of feature maps extracted from HOGs followed bymax-poolinglayer.

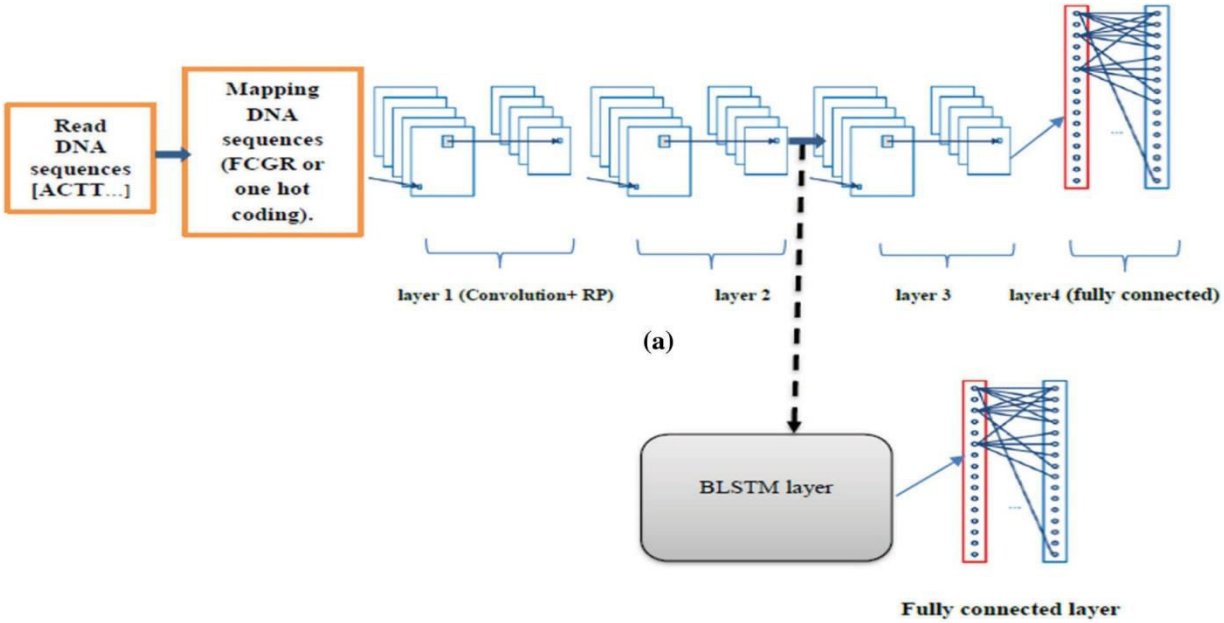


Figure4:Theproposedmodule.(a)CNNbasedonRPmodel,(b)Architectureofthehybridmodel

5 ExperimentalResults

Simulation experiments have been carried out to evaluate the encoded bacterial DNA sequenceclassification based on different approaches for achieving high performance. The DNA sequences havebeen encodedusingtheFCGRalgorithmorbyone-hotcoding.Theparametersused inthesimulationare

the*k*-

mersoftheFCGRalgorithmequalto6.Abatchsizeof128trainingsamplesisemployedtodepicttheperformanceofthehybrid model.Fiveclassificationmodelshavebeenadoptedasfollows:

a) Model1:ClassificationofmappedDNAsequencesusingaclassicalCNNandanRPlayer(sub-samplinglayer).
b) Model 2:Classificationoffeaturemapsextracted fromHOGsusing RNNwithBLSTM.
c) Model3:Classificationof featuremapsextractedwithCNNfollowedbymax-poolingusingRNNwithBLSTM.
d) Model4:ClassificationoffeaturemapsextractedfromCNNfollowedbywaveletpoolingusingRNNwithBLSTM.
e) Model5:ClassificationoffeaturemapsextractedfromCNNfollowedbyRPusingRNNwithBLSTM.

Theproposedmodelshavebeentrainedusing70%oftheinputdataandtestedusingtheremaining30%.A comparison of the accuracy performance among the five models is demonstrated in Tabs. 1–4. Theresultant full-length DNA sequence is specified in Tabs. 1 and 2. Tabs. 3 and 4 are obtained according to500 bp-length sequences. Figs. 5 and 6 show a comparison of the *F*1 score performance among the fivemodels.Accordingtothepreviousresults,theW-CNNfeaturesofBLSTM(model4)havethebest accuracy among all the other models, especially on the genus and family levels. Additionally, the FCGRmapping is more suitable for encoding. Nevertheless, the proposed classical CNN based on RP

consumeslessrunningtime.

Table1:Comparisonbetweenaccuracyscoresformodels(1,2,3,and4)at*k*=6forthefulllength

Classifier	Phylum	Class	Order	Family	Genus
CNNbasedonRP	1	0.9990	0.9910	0.9830	0.9744
HOG featuresbased onRNNwitdBLSTM	1	1	0.9583	0.9400	0.9325
Max-CNN featuresbasedonRNN witdBLSTM	1	1	0.9920	0.9850	0.9735
RP-CNNfeatures based on RNN witd BLSTM	1	1	0.9920	0.9885	0.9835
W-CNNfeaturesbasedonRNNwitdBLSTM	1	1	0.9920	0.9965	0.9950

Table 2: Comparison between accuracy scores for models (1, 2, 3, and 4) using one-hot coding for the fulllength

Classifier	Phylum	Class	Order	Family	Genus
CNNbasedonRP	0.9955	0.9955	0.9340	0.8875	0.8765
HOG featuresbased onRNNwitdBLSTM	0.9950	0.9750	0.9320	0.8800	0.8765
Max-CNN featuresbasedonRNN witdBLSTM	0.9950	0.9945	0.9450	0.9050	0.8975
RP-CNNfeatures based on RNN witd BLSTM	0.9975	0.9955	0.9455	0.9125	0.9025
W-CNN features based on RNNwitdB \LSTM	0.9975	0.9950	0.9500	0.9220	0.9100

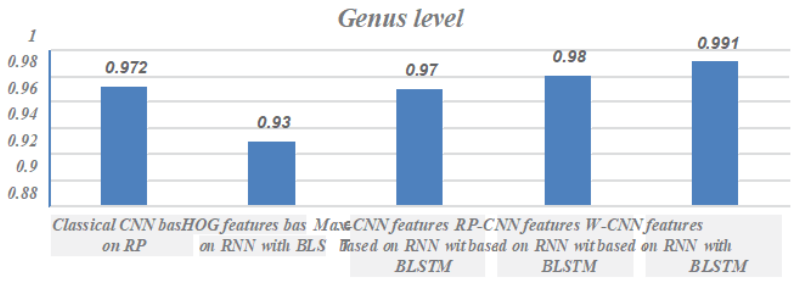
Table3:Comparisonbetweenaccuracyscoresformodels(1,2, 3,and4)at*k*=6for500bp-lengthsequences

Classifier	Phylum	Class	Order	Family	Genus
CNNbasedonRP	0.9960	0.9950	0.9322	0.8356	0.8100
HOG featuresbased onRNNwithBLSTM	0.9960	0.9960	0.9183	0.8340	0.7985
Max-CNN featuresbasedonRNN withBLSTM	0.9960	0.9960	0.9200	0.8365	0.8145
RP-CNNfeatures based on RNN with BLSTM	0.9980	0.9940	0.9365	0.8405	0.8245
W-CNNfeaturesbasedonRNNwithBLSTM	0.9980	0.9950	0.9450	0.8500	0.8295

Table4:Comparisonbetweenaccuracyscoresformodels(1,2,3,and4)usingone-hotcodingfor500bp-

lengthsequences

Classifier	Phylum	Class	Order	Family	Genus
CNNbasedonRP	0.9850	0.9755	0.9040	0.7175	0.7045
HOG featuresbased onRNNwithBLSTM	0.9700	0.9750	0.8920	0.7000	0.6920
Max-CNN featuresbasedonRNN withBLSTM	0.9850	0.9745	0.9050	0.7400	0.7265
RP-CNNfeatures based on RNN with BLSTM	0.9875	0.9755	0.9155	0.7525	0.7375
W-CNNfeaturesbasedonRNNwithBLSTM	0.9880	0.9755	0.9205	0.7625	0.7420



(a)

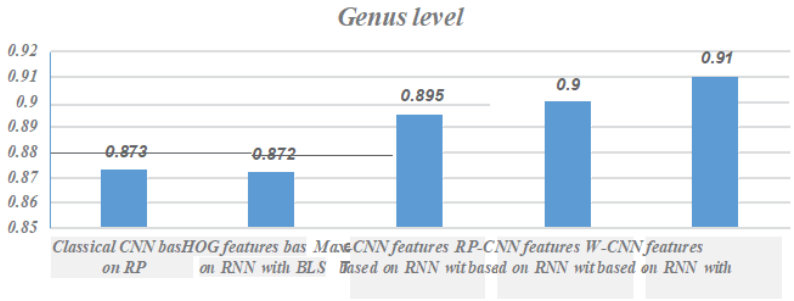


Figure5:Comparisonbetween*F*1scoresformodels(1,2,3,and4)atthegenuslevelforthefulllength. (a)At*k*=6,(b)Usingone-hotcoding6

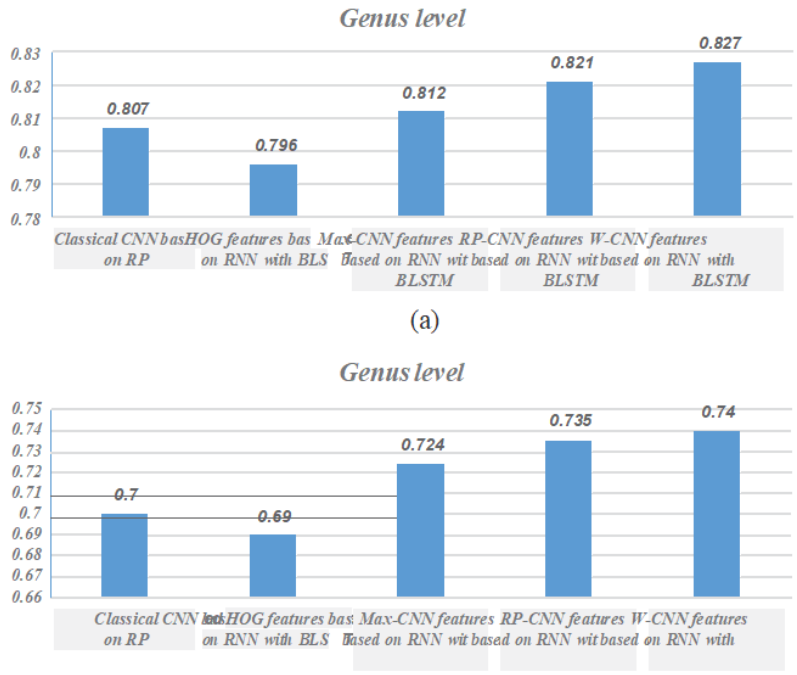


Figure6:Comparisonbetween*F*1scoresformodels(1,2,3,and4)atthegenuslevelfor500bp-lengthsequences.(a)At*k*=6,(b)Usingone-hotcoding

6 Conclusions

In conclusion, this study demonstrates the significant potential of a hybrid deep learning approach for enhancing taxonomic classification, effectively addressing the limitations of traditional methods. By integrating Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), the proposed model successfully leverages both visual and textual data, resulting in improved accuracy and efficiency in species identification. The empirical results indicate that this hybrid framework not only outperforms standalone models but also addresses the complexities inherent in taxonomic classification, particularly when dealing with diverse and intricate datasets. However, challenges related to data quality, model interpretability, and the need for comprehensive training datasets remain. Future research should focus on refining the hybrid model by incorporating advanced techniques such as transfer learning and explainable AI to further enhance performance and usability. Ultimately, this research contributes to the growing body of knowledge in machine learning applications for biodiversity studies, paving the way for more robust and reliable classification systems that can significantly aid in ecological monitoring and conservation efforts.

References

[1] B.Alberts,"*MolecularBiologyoftheCell,*"4thed.,Chapter4.DNAandChromosomes,NewYork:GarlandScience,2002.

[2] D.Moore,"*TheDevelopingGenome:AnIntroductiontoBehavioralEpigenetics,*"UnitedKingdom:OxfordUniversity Press,2015.

[3] B. Tropp andD. Freifelder, "*MolecularBiology: Genes toProteins,*" Chapter4NucleicAcidStructure, 3rd ed.,Sudbury,Mass:JonesandBartlettPublishers,2008.

[4] H.Tettelin,D.Riley,C.CattutoandD.Medini,"Comparativegenomics:Thebacterialpan-genome,"*CurrentOpinioninMicrobiology,*vol.11,no.5,pp.472–477,2008.

[5] HomologyConcepts,[Online].Available:http://en.wikipedia.org/wiki/homology_(biology),lastaccesson11-07-2020.

[6] S.Altschul,W.Gish,W.Miller,E.W.MyersandD.J.Lipman,"Basiclocalalignmentsearchtool,"*J.Mol.Biol,*vol.215,no.3,pp.403–410,1990.

[7] D.LipmanandW.Pearson,"Rapidandsensitiveproteinsimilaritysearches,"*Science,*vol.227,no.4693,pp.1435–1441,1985.

[8] G. BoscoandL.Pinello,"A newfeature selectionmethodologyfork-mers representationof DNAsequences," *CIBB, LNCS, Springer,Heidelberg,*vol.8623, no.4,pp. 99–108, 2015.

[9] G.Bosco,"Alignmentfreedissimilaritiesfornucleosomeclassification," *CIBB,LNCS,Springer,Heidelberg,*vol.9874,no.7,pp.114–128,2016.

[10] S.FernandoandS.Perera,"Empiricalanalysisofdataminingtechniquesforsocialnetwork,"*COMPUSOFT,AnInternationalJournalofAdvancedComputerTechnology,*vol.3,no.2,pp.201–223,2014.

[11] Y.LeCun,Y. BengioandG.Hinton,"Deeplearning,"*Nature,*vol.521, no.2,pp.436–444,2015.

[12] K.SheaandR.Nash,"Anintroductiontoconvolutionalneuralnetworks,"*ArXivPreprintArXiv:1511.08458,*vol.4,pp.1–13,2015.

[13] J. Hochreiter, "Longshort-term memory,"*NeuralComputation,*vol.9,no.8,pp. 1735–1780,1997.

[14] A. Graves, A. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEEInt.Conf.onAcoustics,SpeechandSignalProcessing*,Vancouver,BC,Canada,pp.6645–6649,2013.

[15] T. Mikolov, M. Karaat, L. Burget, J. Cernocky and S. Khudanpur, "Recurrent neural network-based languagemodel,"*InInterspeech,*vol.2,no.6,pp.1045–1048,2010.

[16] R. Wu, S. Yang, D. Leng, Z. Luo and Y. Wang, "Random projected convolutional feature for scene textrecognition," in *Proc. 15th IEEE Int. Conf. on Frontiers in Handwriting Recognition*, Shenzhen, China, pp.132–137,2016.

[17] W. Johnson and J. Lindenstrauss, "Extensions of lipchitz mapping into hilbert space," in *Proc. Conf. in ModernAnalysisandProbability,Amer.Math.Soc.,ofContemporaryMathematics*,Jerusalem,Israel,pp.189–206,1984.

[18] W.El-Shafai,F.Khallaf,E.El-RabaieandF.AbdEl-Samie,"RobustmedicalimageencryptionbasedonDNA-chaos cryptosystem for secure telemedicine and healthcare applications," *Journal of Ambient Intelligence andHumanizedComputing,*vol.3,no.2,pp.1–29,2021.

[19] G.Sakakibara,"ConvolutionalneuralnetworksforclassificationofalignmentsofnonCcodingRNAsequences," *Bioinformatics,*vol.34,no.13,pp.i237–i244,2018.

[20] Y. Wang, K. Hill, S. Singh and L. Kari, "The spectrum of genomic signatures: From dinucleotides to chaos gamerepresentation,"*Gene,*vol.346,no.2,pp.173–185,2005.

[21] R. Rizzo, A. Fiannaca, M. Rosa and A. Urso, "Classification experiments of DNA sequences by using a deepneuralnetworkandchaosgamerepresentation," in*Proc.IEEEInt.Conf.onComputerSystemsandTechnologiesCompSysTech'16*,Palermo,Italy,pp.222–228,2016.

[22] C. Angermueller, T. Pärnamaa, L. Parts and O. Stegle1, "Deep learning for computational biology," *MolecularSystemsBiology,*vol.12,no.7,pp.207–211,2016.

[23] G. Bosco and M. Gangi, "Deep learning architectures for DNA sequence classification," in *Proc. 11th Int.WorkshopofFuzzyLogicandSoftComputingApplications*,Naples,Italy,pp.162–171,2017.

[24] R. Damasevicius, "Analysis of binary feature mapping rules for promoter recognition in imbalanced DNAsequence datasets using support vector machine, intelligent systems," in *Proc. 4th Int. IEEE Conf. IntelligentSystems*,Varna,Bulgaria,pp.11–25,2008.