Computational and Structural Biotechnology Journal 23 (2024) 1631–1640

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/Web server Article

RBProkCNN: Deep learning on appropriate contextual evolutionary information for RNA binding protein discovery in prokaryotes

Upendra Kumar Pradhan [a], Sanchita Naha [b,1], Ritwika Das [c], Ajit Gupta [a], Rajender Parsad [d], Prabina Kumar Meher [a,*,2]

[a] Division of Statistical Genetics, ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India
[b] Division of Computer Applications, ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India
[c] Division of Agricultural Bioinformatics, ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India
[d] ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi 110012, India

ARTICLE INFO

Key words:
RNA-binding proteins
Prediction model
Machine learning
Computational biology
Evolutionary feature

ABSTRACT

RNA-binding proteins (RBPs) are central to key functions such as post-transcriptional regulation, mRNA stability, and adaptation to varied environmental conditions in prokaryotes. While the majority of research has concentrated on eukaryotic RBPs, recent developments underscore the crucial involvement of prokaryotic RBPs. Although computational methods have emerged in recent years to identify RBPs, they have fallen short in accurately identifying prokaryotic RBPs due to their generic nature. To bridge this gap, we introduce RBProkCNN, a novel machine learning-driven computational model meticulously designed for the accurate prediction of prokaryotic RBPs. The prediction process involves the utilization of eight shallow learning algorithms and four deep learning models, incorporating PSSM-based evolutionary features. By leveraging a convolutional neural network (CNN) and evolutionarily significant features selected through extreme gradient boosting variable importance measure, RBProkCNN achieved the highest accuracy in five-fold cross-validation, yielding 98.04% auROC and 98.19% auPRC. Furthermore, RBProkCNN demonstrated robust performance with an independent dataset, showcasing a commendable 95.77% auROC and 95.78% auPRC. Noteworthy is its superior predictive accuracy when compared to several state-of-the-art existing models. RBProkCNN is available as an online prediction tool (https://iasri-sg.icar.gov.in/rbprokcnn/), offering free access to interested users. This tool represents a substantial contribution, enriching the array of resources available for the accurate and efficient prediction of prokaryotic RBPs.

1. Introduction

Ubiquitous across all living organisms, RNA-binding proteins (RBPs) play crucial roles for regulating a wide range of cellular functions including the regulation of post-transcriptional genes [1–5]. While the historical emphasis has been on eukaryotic RBPs, recent investigations have brought to light the critical significance of their prokaryotic counterparts [6–8]. By acting as a regulator of various cellular processes and playing crucial role in sculpting the dynamic landscape of bacterial gene expression, prokaryotic RBPs contribute significantly to the adaptation of prokaryotes across diverse environmental conditions [9,10]. A noteworthy aspect of prokaryotic RBPs further lies in their participation in governing bacterial virulence and pathogenicity, pinpointing potential targets for novel antibiotic development [11–15]. In sum, identification of RBPs in prokaryotes holds significant importance for deciphering the gene expression patterns and propelling the field of precision medicine forward. Traditionally, wet lab techniques have been used to identify RNA-protein interactions. However, wet lab experiments encounter challenges due to their expensive nature and complex biotechnological requirements [16,17], albeit proficient in accurately identifying RBPs. Thus, there is a growing demand for computational methods capable of proteome-wide prediction of RBPs, which warrants further experimental investigation.

The existing approaches for identification of RBPs falls under two

* Corresponding author.
E-mail address: prabina.meher@icar.gov.in (P.K. Meher).
[1] Joint first author.
[2] ORCID: https://orcid.org/0000–0002-7098–8785

https://doi.org/10.1016/j.csbj.2024.04.034
Received 16 February 2024; Received in revised form 12 April 2024; Accepted 12 April 2024
Available online 15 April 2024

major categories: (i) machine learning-based techniques and (ii) template-based techniques. Template-based techniques gauge the resemblance of a query protein to a template RNA-binding domain (RBD) or RBP to discern RBPs [18–21]. However, the pre-established RBDs are absent in nearly half of experimentally identified RBPs [5,22], and also the presence of RBDs in proteins does not always correlate with them being RBPs [3], implying ineffectiveness of the template-based methods. Conversely, machine learning-based methods have gained prominence in predicting RBPs that uses annotated datasets encompassing both RBPs and non-RBPs to train learning algorithms.

For training of machine learning models, features derived from the protein sequences as well as from the three-dimensional structures have been employed in existing studies. Though some models like BindUp [23], NucleicNet [24], and NAbind [25] have utilized three-dimensional structural features, the algorithms currently in use have predominantly utilized sequence-derived features due to ease of getting sequence data compared to three-dimensional structure data. For instance, RNApred [26] employed support vector machine (SVM) coupled with position-specific scoring matrix (PSSM)-derived features to differentiate RBPs from non-RBPs. Zhang and Liu [27] devised RBPPred, for predicting RBPs with SVM, incorporating physico-chemical properties and PSSM-derived features of the protein sequences. By integrating the sequence-derived information, an ensemble learning model named iDRBP-EL was developed by Wang et al. [28], for the discovery of RBPs and DNA-binding proteins (DBPs). Another computational model called IDRBP-PPCT was introduced by Wang et al. [29], where the random forest method was used for prediction using a novel feature representation technique, referred as Position-Specific Frequency Matrix and Cross Transformation (PPCT). Feng et al. [30] developed iDRBP-ECHF for identifying DBPs and RBPs based on extensible cubic hybrid framework. A plant-specific RBP prediction model, RBPLight was developed by Pradhan et al. [31] using evolutionary features derived from PSSM profiles and light gradient boosting method (LightGBM).

In addition to shallow learning algorithms, deep learning frameworks have also been employed for predicting RBPs. Zheng et al. [32] employed convolutional neural network (CNN) by incorporating the protein feature of RBPPred to develop Deep-RBPPred. To address cross-prediction challenges, Zhang et al. [33] established iDRBP_MMC, a multi-label learning model based on CNN, for predicting DBPs and RBPs. Fusing CNN with LSTM, Zhang et al. [34] devised a two-level computational model for identification of DBPs and RBPs. In another investigation, Zhang et al. [35] developed PreRBP-TL computational model to detect species-specific RBPs, leveraging transfer learning. The RBP-TSTL [36] is another RBP prediction model which was developed by integrating knowledge from pre-trained RBP datasets and from a self-supervised pre-trained model. Another model iDRPro-SC was recently introduced by Yan et al. [37], for prediction of both DBPs and RBPs based on sequence information using an ensemble learning technique.

Most of the above stated learning models have been utilized the RBP data from diverse eukaryotic species and a limited number of prokaryotes, resulting in generic models. However, RBPs exhibit specificity not only to individual species but also to lineage-specific families [6,22,31,38]. Therefore, the current generic models may lack the information required for accurate prediction of prokaryote-specific RBPs. While some models like PreRBP-TL [35] and RBP-TSTL [36] have been tested on certain bacteria species, their accuracy in predicting RBPs for other prokaryotic species remains suboptimal. In other words, despite the notable progress in RBP prediction, the development of models specific to prokaryotes has been largely overlooked. Hence, there is a pressing need to devise computational methods tailored for predicting prokaryote-specific RBPs. Here, we proposed a new computational model, RBProkCNN, designed explicitly for predicting prokaryote-specific RBPs.

2. Material and methods

2.1. Retrieval and processing of sequence data

Prokaryotes are broadly categorized into two taxonomic groups, bacteria and archaea. Protein sequences corresponding to bacteria (taxonomy id: 2) and archaea (taxonomy id: 2157) were sourced from the UniProt database [39] as of June 16, 2023. These sequences were utilized to construct datasets for RBPs (positive) and non-RBPs (negative). The protein sequences that were confirmed to be annotated with the Gene Ontology (GO) term "RNA-binding" (GO: 0003723 and it's child terms) were defined as RBPs, while proteins lacking this annotation were considered as non-RBPs. A total of 48,626 RBP sequences and 2,80,033 non-RBP sequences available for 741 prokaryotic species were retrieved. Sequences with fewer than 50 amino acids as well as having non-standard residues were excluded from the analysis. To eliminate accuracy bias arising from the presence of homologous sequences, the CD-HIT algorithm [40] was applied to discard sequences with $>$ 25% sequence identity to any other sequences in both positive and negative datasets. After processing, a set of 1480 RBP sequences and 29,971 non-RBP sequences were retained for the analysis.

2.2. Training and independent test datasets construction

Out of the initial pool of 1480 RBP sequences, approximately 20%, totalling 280 sequences, were randomly chosen and set aside to form the positive independent test set. The remaining 1200 sequences, constituting roughly 80%, were utilized as training set for the RBP class (positive). In order to minimize the biased effect of major class (class having larger number of instances) on prediction performance, a balanced training dataset was prepared with same number of instances of RBP and non-RBP (negative) classes. In other words, the training dataset was composed of an equal number of positive and negative instances, with 1200 negative instances randomly selected from the pool of 29,971 non-RBP sequences. From the remaining 17,971 non-RBP sequences, 280 sequences were randomly chosen to construct the negative independent test set. In summary, the training dataset comprised of 1200 RBP and 1200 non-RBP sequences, while the independent test set included 280 RBP and 280 non-RBP sequences.

2.3. Evolutionary feature generation

The PSSM profile for each protein sequence was generated by running PSI-BLAST [41] on the non-redundant database NRDB90 [42] with e-value 0.0001. If we let $P=\left(\left(p_{i,j}\right)\right)$ be the PSSM matrix representation for a given protein sequence of L amino acids long, P will be of dimension $L \times 20$. Using the PSSM profile of each sequence, we generated k-separated bi-gram features (KBGM) and tri-gram features (TRGM), respectively called as KBGM_PSSM and TRGM_PSSM features. The mathematical derivation of both feature set is as follows:

In a protein sequence of L amino acids long, the KBGM_PSSM feature $f_{m,n}(k)$ for the amino acid pair (m,n) with distance k, can be computed as $f_{m,n}(k)=\sum_{i=1}^{L-k} p_{i,m}p_{i+k,n}; 1 \le m,n \le 20$, where $p_{i,m}$ is the scoring value of the PSSM corresponding to i^{th} position and the amino acid m. Since the total possible combination of amino acid pair (m,n) is 400, there will be 400 elements representing 400 amino acid transitions in the resultant feature vector such as $\left\{f_{1,1}(k), f_{1,2}(k), \ldots, f_{1,20}(k), f_{2,1}(k), \ldots, f_{2,20}(k), \ldots, f_{20,1}(k), \ldots, f_{20,20}(k)\right\}$. In this study, we considered the value of $k=2$. Using the same PSSM profile, the TRGM_PSSM feature $f(m,n,r)$ corresponding to the amino acid trio (m,n,r) can be computed as $f(m,n,r)=\sum_{i=1}^{L-2} P_{i,m}P_{i+1,n}P_{i+2,r}; 1 \le m,n,r \le 20$. Since three are 8000 possible combinations of the amino acid trio (m,n,r), there will be 8000 element in the resultant feature vector.

2.4. Prediction algorithms and feature selection

Initially, we evaluated the performance of 8 shallow learning models for prediction of prokaryotic RBPs. Then, the best performing shallow learning models were compared with that of 4 deep learning models. The details of the software used and the parameter setup for implementing these learning models are shown in Table 1. Feature selection strategy was also applied to alleviate computational load and enhance classification accuracy by eliminating redundant and irrelevant features [55]. Two feature ranking algorithms such as extreme gradient boosting variable importance measure (XGB-VIM) [56] and light gradient boosting machine variable importance measure (LGBM-VIM) [46], were utilized to select significant and pertinent features. The performance of the selected learning algorithms was then evaluated using the top-ranked features. Though there are several feature selection techniques available, we chose to focus on XGB-VIM and LGBM-VIM because (i) XGB-VIM and LGBM-VIM are the feature selection models based on gradient boosting algorithms and are known for their robustness in identifying relevant features while reducing the effects of noise and irrelevant variables, (ii) by focusing on variables that contribute most to model performance, XGB-VIM and LGBM-VIM help improve model generalization and reduce over fitting and, (iii) XGB-VIM and LGBM-VIM algorithms are computationally efficient, suitable for handling large datasets with high-dimensional feature spaces (in the present case >8000 features).

Table 1
Parameter configuration and software utilized for implementation of both shallow and deep learning models.

ML models	Parameter setting	Software used
Support Vector Machine (SVM) [43]	kernel = "rbf", $\gamma = 1/$ #column, cost = 1	*e1071* R-package
Extreme Gradient Boosting (XGBoost) [44]	max_depth = 3, $\eta = 1$, nrounds = 2, objective = "logistic"	*xgboost* R-package
Random Forest (RF) [45]	ntree = 1000, mtry = $\sqrt{\#\text{column}}$	*randomForest* R-package
Light Gradient Boosting (LightGBM) [46]	objective= 'binary', boosting= "gbdt", learning_rate = 0.1, num_leaves = 31, nrounds = 1000	*lightgbm* R-package
Stochastic Gradient Descent(SGD) [47]	learning rate = 0.01, max_iter= 1000, batch_size= 32, tol= 1e-3	*scikit-learn* Python library
Gradient Boosting Decision Tree (GBDT) [48]	n_estimators= 1000, learning_rate= 0.01, max_depth= 3	*scikit-learn* Python library
Adaptive Boosting (AdaBoost) [49]	v = 5, mfinal = 1000	*adabag* R-package
Bagging [50]	nbagg= 25	*ipred* R-package
Convolutional Neural Networks (CNN) [51]	1st Conv1D: 5 × 1, 5 kernels; 1st Maxpooling1D: 2 × 1; 2nd Conv1D: 5 × 1, 10 kernels; 2nd Maxpooling1D: 2 × 1; ReLU activation function; Dense layer neurons = 500; 2 dropout layers with rate 0.2, Adam optimizer, Softmax loss function, Binary cross-entropy, epoch = 100, batch size = 20, learning rate = 0.001	*TensorFlow* python module
Long Short-Term Memory (LSTM) [52]	LSTM units = 64, time step = 10, with dropout = 0.2; Dense layer neurons = 500; ReLU activation function; 2 dropout layers with rate 0.5, Adam optimizer, Softmax loss function, Binary cross-entropy, epoch = 100, batch size = 20, learning rate = 0.001	*PyTorch* module of python
Bidirectional LSTM (Bi-LSTM) [53]	LSTM units = 64, time step = 10, with dropout = 0.2; merge-mode = "average"; Dense layer neurons = 500; ReLU activation function; 2 dropout layers with rate 0.5, Adam optimizer, Softmax loss function, Binary cross-entropy, epoch = 100, batch size = 20, learning rate = 0.001	*PyTorch* module of python
Gated Recurrent Unit (GRU) [54]	GRU units = 64, time step = 10, with dropout = 0.2 and recurrent dropout = 0.2; Dense layer neurons = 500; ReLU activation function; 2 dropout layers with rate 0.5, Adam optimizer, Softmax loss function, Binary cross-entropy, epoch = 100, batch size = 20, learning rate = 0.001	*TensorFlow* python module

2.5. Evaluating the performance through cross-validation

We followed a 5-fold cross-validation procedure to evaluate the performance of the learning algorithms, where the metrics of the learning algorithms were obtained by averaging their performance across the five validation sets of the cross-validation. Following performance metrics were employed for quantitative assessment of the performance of the learning models.

$$Accuracy = \frac{T^+ + T^-}{T^+ + F^- + T^- + F^+}$$

$$Precision = \frac{T^+}{T^+ + F^+}$$

$$F1 - Score = \frac{2T^+}{2T^+ + F^+ + F^-}$$

$$Matthews\ \ Correlation\ \ Coefficient(MCC) = \frac{(T^+ \times T^-) - (F^+ \times F^-)}{\sqrt{(T^+ + F^+)(T^+ + F^-)(T^- + F^+)(T^- + F^-)}}$$

$$Area\ \ under\ \ receiver\ \ operating\ \ characteristic\ \ curve(auROC) = \int_0^1 \frac{T^+}{F^- + T^+} d\left(\frac{F^+}{T^- + F^+}\right)$$

$$Area\ \ under\ \ precision - recall\ \ curve(auPRC) = \int_0^1 \frac{T^+}{T^+ + F^+} d\left(\frac{T^+}{F^- + T^+}\right)$$

In the above defined metrics, T^+, T^-, F^+ and F^- stands for the true positive, true negative, false positive and false negative, respectively. A schematic flowchart illustrating each step of the proposed approach is presented in Fig. 1.

3. Results

3.1. Analysis of shallow learning algorithms and evolutionary features

Performance of 8 shallow learning models were evaluated with 15 different PSSM-derived feature sets. The LightGBM, RF, SVM and XGBoost models achieved higher accuracy across the feature sets, as compared to the other learning models. As far as feature set is concerned, higher accuracies were observed for TRGM_PSSM and KBGM_PSSM feature sets for most of the learning algorithms. The highest auROC (97.99%) was achieved by SVM with TRGM_PSSM features, followed by RF (97.71%) with KBGM_PSSM features (Fig. 2.). Similarly, SVM achieved the highest auPRC (98.26%) with TRGM_PSSM, followed by RF (98.01%) with KBGM_PSSM features. If the performance metrics were averaged across the feature sets, the highest accuracy (89.45%), auROC (95.49%) and auPRC (95.34%) were observed with the LightGBM followed by XGBoost (accuracy: 88.81%, auROC: 95.10%, auPRC: 94.89%) (Fig. 2.). Similarly, while performance metrics were averaged across the learning algorithms, the highest accuracy (90.11%), auROC (95.73%) and auPRC (95.62%) were obtained with KBGM_PSSM features (Fig. 2.). Taking all the performance metrics into account, the TRGM_PSSM and KBGM_PSSM features were selected for further

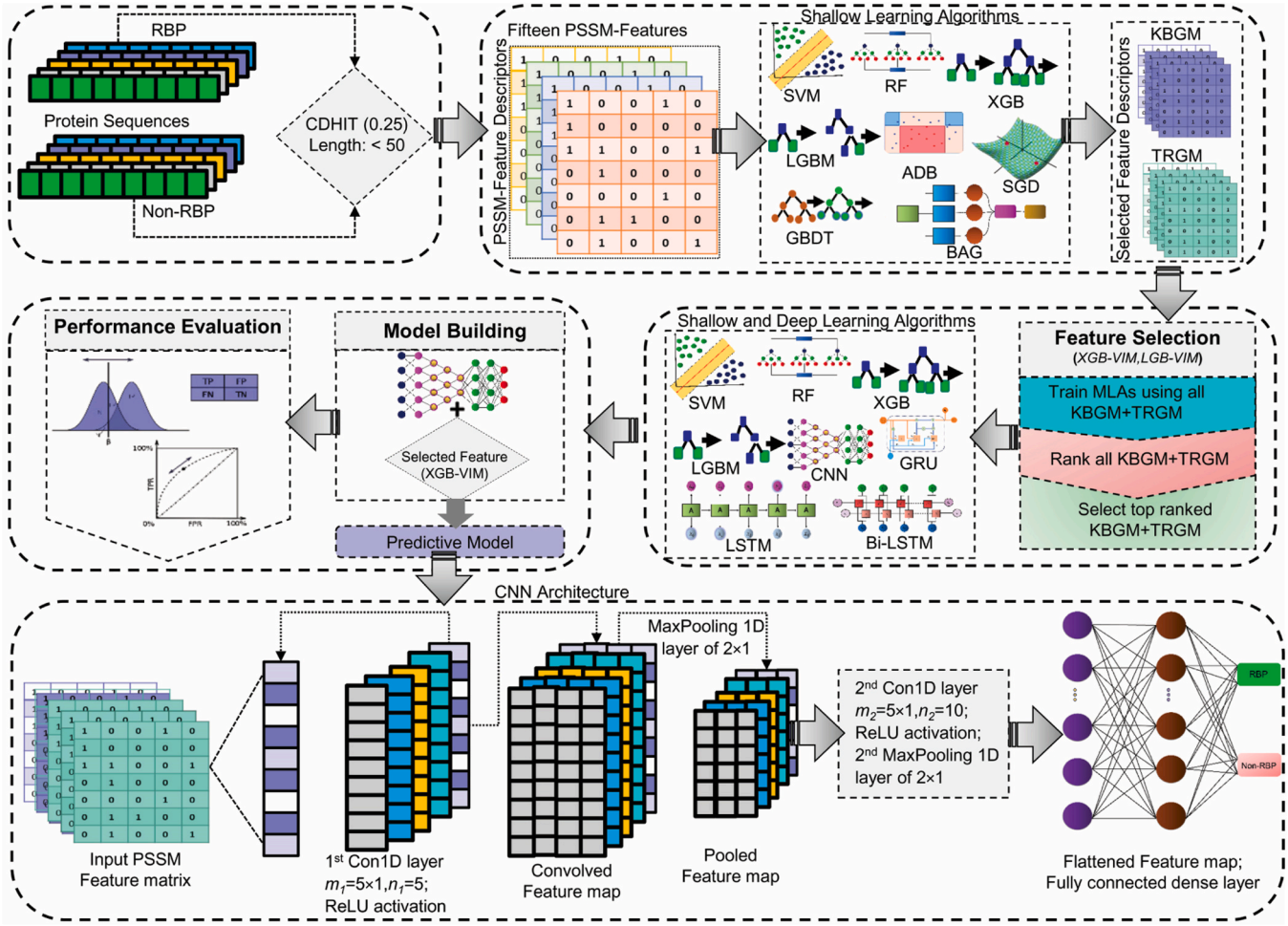

Fig. 1. Flow diagram depicts the different steps followed to develop the proposed RBP prediction model.

analysis with the four best performing models such as LightGBM, RF, SVM and XGBoost.

3.2. Evaluation of deep learning models for predicting RBPs

The performance of the 4 best learning models (SVM, RF, LightGBM and XGBoost) were compared with that of 4 deep learning models (LSTM, Bi-LSTM, CNN and GRU) using the KBGM_PSSM, TRGM_PSSM and KBGM+TRGM_PSSM features. It was observed that performance metrics enhanced while combined features (KBGM+TRGM_PSSM) were employed, as compared to the individual feature sets (KBGM_PSSM, TRGM_PSSM), for almost all the machine learning models (Fig. 3.). All the 4 deep learning models achieved > 90% accuracy, > 95% auROC and auPRC for all the three feature sets (Fig. 3.). Among the deep learning models, GRU achieved the highest accuracy (94.66%) and auROC (98.28%) with TRGM_PSSM features and the highest auPRC (98.28%) with KBGM+TRGM_PSSM features (Fig. 3.). It was also observed that the deep learning model GRU achived a little higher auROC (98.24%) and auPRC (98.28%) as compared to the best performing shallow learning model RF (auROC: 97.96%, auPRC: 98.25%) (Fig. 3.).

3.3. Feature selection analysis

Though higher performance metrics were achieved with KBGM+TRGM_PSSM features, the feature dimension is very high (8400 features) which may lead to over prediction. Thus, variable importance measure of XGBoost and LightGBM were utilized for selection of relevant and non-redundant predictor variables. We selected the top 1400 features and performance metrics (auROC and auPRC) were computed through 5-fold cross-validation by accounting 10 selected features sequentially. It was seen that CNN achieved the higher performance metrics followed by LightGBM across the selected features, though GRU was seen to be the best performer with full 8400 features (Fig. 4.). In particular, CNN achieved the highest performance metrics with 920 XGB-VIM selected features (auROC: 98.04%, auPRC: 98.19%) and 1320 LGBM-VIM selected features (auROC: 97.99%, auPRC: 98.21%) (Fig. 4).

3.4. Prediction analysis for independent test set

The CNN model trained with 920 XGB-VIM and 1320 LGBM-VIM selected features was further employed for prediction of the independent dataset which comprises 280 sequences from both RBP and non-RBP classes. The accuracy of the independent dataset was found higher with the XGB-VIM selected features (90.45%) as compared to the LGBM-VIM selected features (89.76%) (Table 2). The auROC and auPRC with XGB-VIM features (95.77%, 95.78%) was found a little higher than that of LGBM-VIM selected features (95.07%, 95.38%) (Table 2). The precision with XGB-VIM features (91.82%) was found ~4% higher than compared to the LGBM-VIM features (87.97%) (Table 2).

3.5. Comparison with existing models

Using the independent dataset, performance of the proposed approach (CNN with 920 XGB-VIM selected features) was further

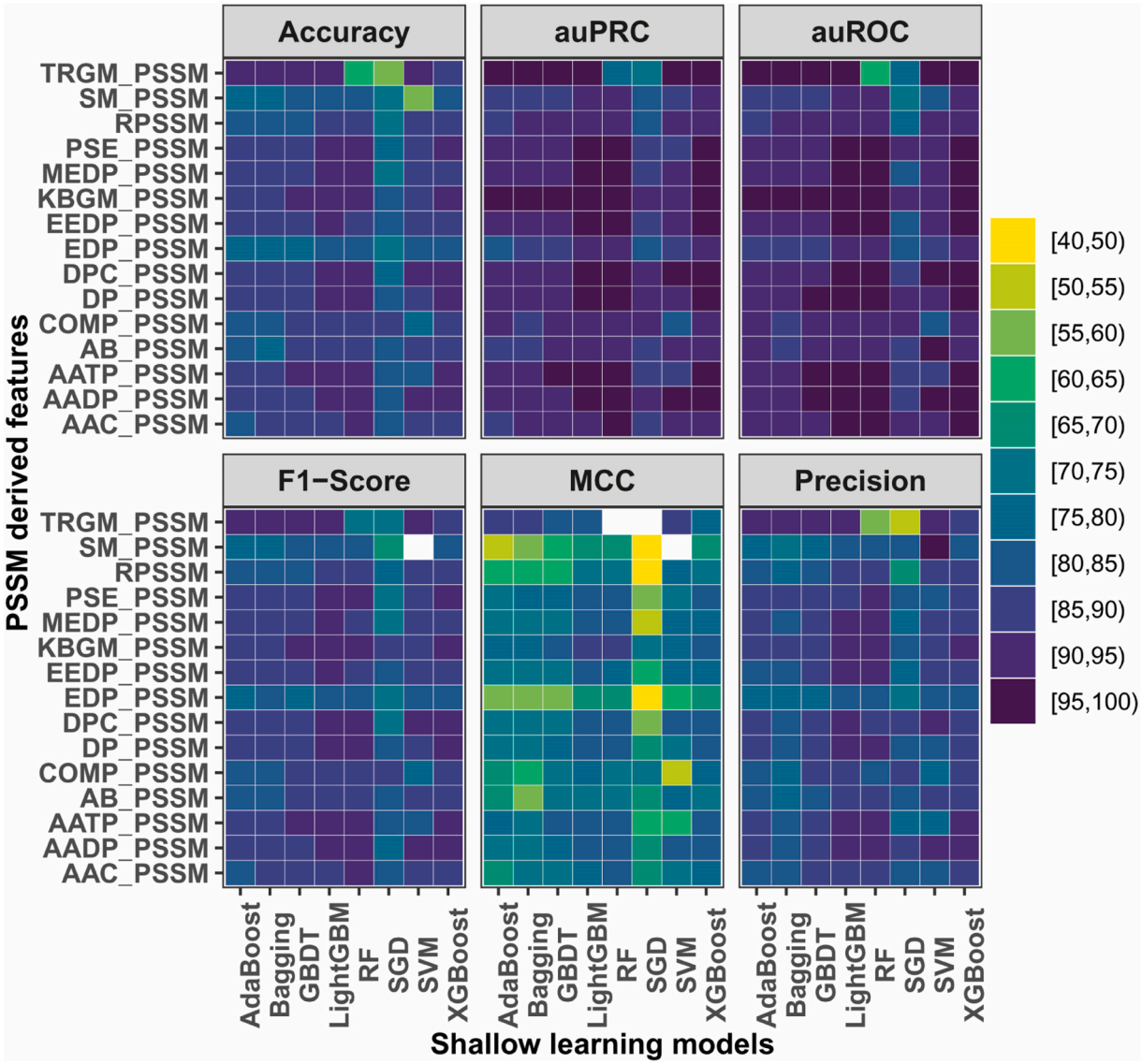

Fig. 2. Heatmaps of the performance metrics of different shallow learning algorithms while prediction was independently done with individual PSSM-derived features.

compared with 10 existing prokaryotic RBP prediction models. None of the existing models were able to achieve > 90% accuracy. Among the existing models, PreRBP-TL achieved the highest performance metrics (Table 3). However, the accuracy of the proposed model (90.45%) was found ~1% higher than that of PreRBP-TL (89.43%) (Table 3). Though auROC of the proposed model (95.77) was found at par with that of PreRBP-TL (95.44%), the auPRC (95.78) was found ~2% higher than that of PreRBP-TL (93.84%) (Table 3). Similarly, the F1-Score of the proposed model (89.98%) and PreRBP-TL (89.28) was found at par, but the proposed model (80.93%) achieved ~2% higher MCC than that of preRBP-TL (78.90%) (Table 3).

3.6. Comparative analysis using the training dataset of PreRBP-TL

From Table 3, it was observed that the PreRBP-TL achieved higher accuracy among the existing RBP prediction models. Thus, to further make a comparison with PreRBP-TL, the performance of RBProkCNN was evaluated using the training dataset of PreRBP-TL. Since the cross-validation accuracy of PreRBP-TL have been reported for two species (*E. coli* and *Salmonella*), the performance of RBProkCNN was evaluated using the training dataset of *E. coli* (351 RBPs and 2819 nonRBPs) and *Salmonella* (206 RBPs and 1107 nonRBPs) following five-fold cross-validation and compared with that of PreRBP-TL. It was found that for *E. coli*, auROC of RBProkCNN (93.62%) achieved a little higher accuracy than that of PreRBP-TL (93.33%), whereas for *Salmonella*, PreRBP-TL achieved higher auROC (95.20%) than that of RBProkCNN (93.87%) (Table 4). As far as auPRC is concerned, RBProkCNN achieved better accuracy as compared to the PreRBP-TL, for both *E.coli* and *Salmonella* (Table 4). Since the dataset is imbalanced, auPRC is a better measure than auROC as it takes into account the information of both positive and negative classes. Thus, on the basis of auPRC measure, it can be said that the proposed model may achieve higher accuracy than that of PreRBP-TL.

3.7. Comparison with PreRBP-TL using another independent dataset

The performance advantage of RBProkCNN over PreRBP-TL was observed to be small, which may be due the initial large language generative model of PreRBP-TL where the test data set of RBProkCNN were included in the model training. Nonetheless, to further validate the higher performance of RBProkCNN over PreRBP-TL, we prepared another independent test dataset comprising prokaryotic RBP sequences

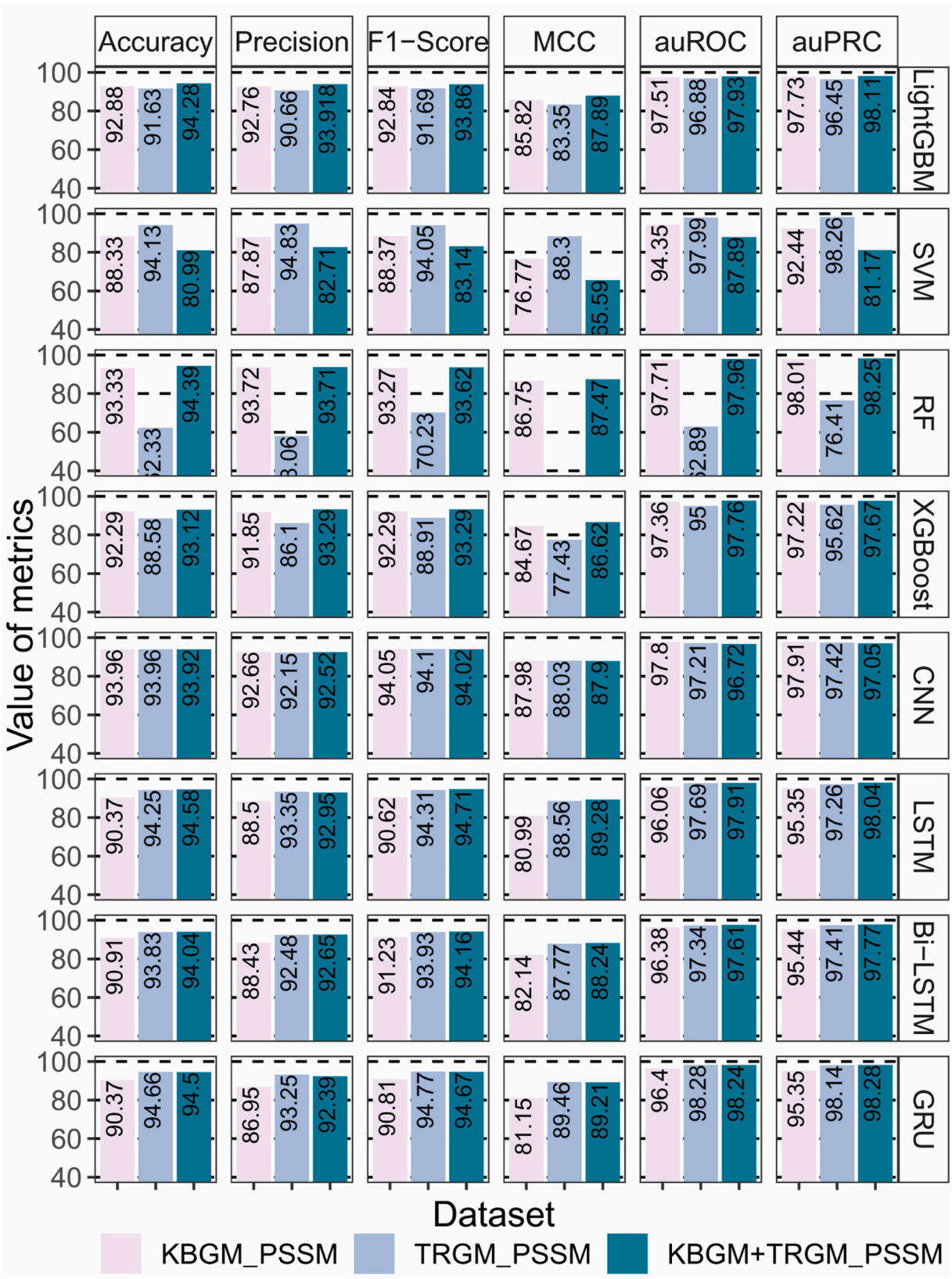

Fig. 3. Bar plots of the performance metrics of the shallow learning and deep learning models with KBGM_PSSM, TRGM_PSSM and KBGM+TRGM_PSSM features.

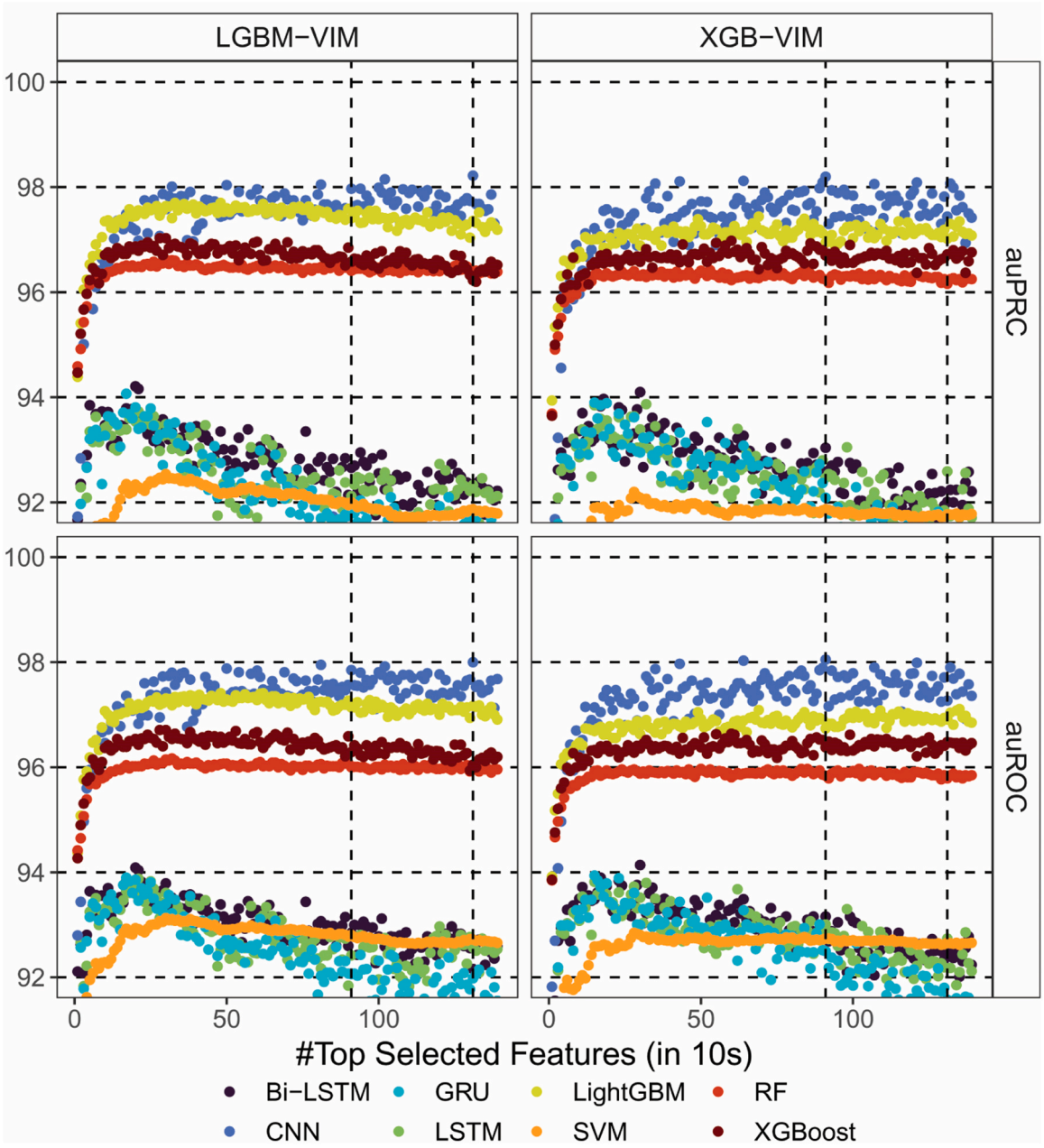

Fig. 4. Dot plots of the auROC and auPRC of the machine learning models while prediction was made using the XGB-VIM and LGBM-VIM selected features.

Table 2
Performance metrics of the CNN on independent dataset while the prediction was performed using 920 XGB-VIM and 1320 LGVM-VIM selected features.

Feature	Accuracy	Precision	F1-Score	MCC	auROC	auPRC
XGB-VIM (920)	90.45	91.82	89.98	80.93	95.77	95.78
LGBM-VIM (1320)	89.76	87.97	89.67	79.58	95.07	95.38

of four different species such as *Haemophilus Influenzae (116), Mycobacterium tuberculosis (97), Shigella flexneri (108)* and *Yersinia pestis (96)*. It was ensured that none of these species were present in the training set of RBProkCNN as well as PreRBP-TL. From Table 5, it was observed that RBProkCNN achieved higher accuracy as compared to the PreRBP-TL model, for all the four species.

4. Discussion

The establishment of a suitable dataset is pivotal in machine learning-based classification tasks, especially in the context of computational biology and bioinformatics [57]. In other words, utilization of a dataset encompassing a larger number of species holds the potential to capture a broader range of conserved patterns in prokaryotic species. For RBP prediction task, the existing models utilised protein sequences from both eukaryotic and prokaryotic sources [26–30,32–34,36,37], where the eukaryotic dataset's size significantly exceeds that of prokaryotic species. This likely contributes to the lower accuracy observed in prokaryotic-specific RBP prediction in existing tools. To enhance generalizability across diverse prokaryotic species and capture universal features among them, our study incorporates 741 prokaryotic species, a substantial increase compared to the fewer than 100 prokaryotic species considered in earlier studies.

The RBPs and non-RBPs were subjected for the removal of highly

Table 3
Comparison of the performance metrics of the proposed approach with that of existing RBP prediction models, using an independent test dataset.

Models	Accuracy	Precision	F1-Sscore	MCC	auROC	auPRC
RNApred	69.15	64.33	72.01	40.12	-	-
RBPPred	68.80	73.36	63.56	38.16	78.98	73.13
Deep-RBPPred (Balanced)	60.83	57.46	64.80	23.19	66.11	64.57
Deep-RBPPred (Unbalanced)	64.12	64.20	61.45	28.11	70.21	68.31
iDRBP_MMC	84.75	96.15	81.97	71.55	-	-
DeepDRBP-2 L	87.18	85.03	87.11	74.45	-	-
iDRBP-EL	66.90	73.06	59.62	34.80	-	-
iDRBP-ECHF	54.25	76.67	14.84	13.19	-	-
IDRBP-PPCT	78.68	77.16	78.38	57.40	-	-
PreRBP-TL (*E. coli*)	89.43	87.89	89.28	78.90	95.21	92.59
PreRBP-TL (*Salmonella*)	88.73	91.19	87.99	77.58	95.44	92.44
PreRBP-TL (*Bacillus*)	85.27	90.46	83.69	71.05	95.03	93.84
PreRBP-TL (*Staphylococcus*)	84.75	89.67	83.14	69.97	93.55	89.62
iDRPro-SC	58.40	59.43	51.21	16.63	-	-
Proposed	90.45	91.82	89.98	80.93	95.77	95.78

The auROc and auPRC could not be computed for some tools due to the unavailability of probability of prediction.

Table 4
Performance comparison of PreRBP-TL and the RBProkCNN using the training datasets of PreRBP-TL.

Model	Species	Accuracy	Precision	F1-Score	MCC	auROC	auPRC
RBProkCNN	*E. coli*	87.54	84.21	80.32	78.14	93.62	83.57
	Salmonella	87.35	81.11	79.54	75.86	93.87	85.19
PreRBP-TL	*E. coli*	-	-	-	-	93.33	81.25
	Salmonella	-	-	-	-	95.20	84.52

Only auROC and auPRC have been reported in PreRBP-TL paper

Table 5
Species-specific prediction accuracy of PreRBP-TL and the proposed model RBProkCNN.

Species	PreRBP-TL				RBProkCNN
	E. coli	*Salmonella*	*Bacillus*	*Streptococcus*	
Haemophilus Influenzae	91.38	87.07	76.72	63.79	93.10
Mycobacterium tuberculosis	88.66	81.44	77.32	69.07	91.75
Shigella flexneri	93.52	91.67	79.63	65.74	94.44
Yersinia pestis	92.71	88.54	81.25	70.83	98.96

The results show the percentage of correctly predicted RBPs among all the RBP sequences

similar sequences using the CD-HIT program because the presence of highly similar sequences introduce homologous bias in the prediction accuracy which in turn leads to the overestimation of the prediction accuracy. Another advantage of using CD-HIT is reduction in computational complexity due to the removal of redundant sequences. There is no thumb rule to select the sequence identity threshold. In this study, we preferred to set the threshold 0.25 which means the sequences that shared more than 25% similarity with other sequences gets eliminated. In several existing studies, the threshold has been set to 0.4. However, a higher threshold will result in less stringent clustering, allowing for more sequences to be grouped together, even if they are not more similar. This can lead to a loss of sequence diversity within clusters, potentially overlooking important variations or subtypes present in the dataset. To study the effects of different thresholds, the prediction analysis was performed using the training datasets at 3 different sequence identity threshold such as 0.3, 0.35 and 0.4. It was observed that with increase in the sequence identity threshold, the prediction accuracies were increased (Supplementary Fig. 1).

Given the limited availability of structural information for RBPs, the majority of existing prediction models relied on sequence-derived features and employed machine learning algorithms for prediction. Notably, among sequence-based features, those derived from PSSM profiles of protein sequences have demonstrated effectiveness in various studies [26–37]. This improvement may stem from the capability of PSSM-based features to capture context-dependent information, conservation patterns, and evolutionary insights. Though we initially considered fifteen distinct PSSM-based evolutionary features, two PSSM-derived feature sets, namely KBGM_PSSM and TRGM_PSSM, were selected to harness the complementary information provided by PSSM and other sequence-derived features. While previous RBP prediction models have also incorporated PSSM-derived features like PSSM-400 [27,32,58,59], BLOSUM62 [28,29,33–35], and PSSM-TPC [31,60], the features KBGM_PSSM and TRGM_PSSM have not been explored for RBP prediction in earlier studies.

In addition to using of KBGM_PSSM and TRGM_PSSM feature set independently, performance of the learning algorithms were also evaluated with combined feature set (8400 features). While ensemble of multiple features effectively encodes protein sequence data, it may introduce redundancy and noise, potentially diminishing the model's effectiveness. Therefore, selection of relevant and non-redundant features is essential. Among existing RBP prediction models, only the RBPpred model [27] utilized the minimum redundant maximum relevant criteria (mRmR) [61] for feature selection. For the current study, we employed two different feature selection techniques, LGBM-VIM [46] and XGB-VIM [56], to choose pertinent and non-redundant features. It was observed that CNN achieved higher performance accuracy with 920 XGB-VIM selected features, although GRU demonstrated the highest accuracy with the full 8400 features. Therefore, CNN with 920 XGB-VIM selected features was chosen for developing the RBProkCNN model. The CNN model has also been identified as the most successful algorithm for RBP prediction task [32,33].

There was an obvious imbalance in the positive and negative data, with larger size for the negative set. However, we utilized the balanced dataset for prediction analysis that comprises equal instances from both positive and negative sets, where the instances of the negative set were sampled from the whole negative dataset. Using of balanced dataset has

certain advantages over the imbalanced dataset that (i) it ensures that the model is trained on a representative sample of each class, preventing bias towards the majority class, (ii) machine learning algorithms often perform better with balanced datasets and (iii) appropriate for evaluating the performance of the model across all classes, as each class has an equal contribution to the evaluation metrics. Nonetheless, the performance of RBProkCNN was also evaluated using the imbalanced training datasets of best performing existing model PreRBP-TL and was found to perform better than that of PreRBP-TL. Thus, it may be said that the RBProkCNN may perform better both with balanced and imbalanced datasets.

The performance RBProkCNN was also assessed on a blind test dataset to demonstrate its robustness and generalization ability, where the performance metrics were found to be closer to that of cross-validation. To further validate the model's reliability, we compared RBProkCNN's performance with existing state-of-the-art methods using the same blind test dataset. The RBProkCNN exhibited higher accuracy compared to the other available methods. In other words, the existing methods were observed achieving less accuracy for predicting prokaryotic-specific RBPs than that of eukaryotic species like humans and mice. The RBPs are known to be highly tissue- and lineage-specific [31,38,62,63], whereas the existing models have been developed based on protein sequences of several eukaryotic and few prokaryotic species, making them less accurate for predicting prokaryote-specific RBPs. Thus, RBProkCNN is believed to address this gap by providing a model tailored for predicting prokaryote-specific RBPs, underscoring the importance of accurate predictions in unravelling gene expression patterns and enhancing our understanding of prokaryotic gene regulation.

5. Conclusion

RNA-binding proteins (RBPs) are indispensable elements across diverse life forms, influencing crucial cellular functions. While the majority of research has centered on eukaryotic RBPs, recent insights underscore the pivotal role of prokaryotic RBPs in bacterial gene regulation, virulence, and adaptation to environmental conditions. The present study introduces RBProkCNN (https://iasri-sg.icar.gov.in/rbprokcnn/), a novel computational tool specifically tailored for predicting prokaryote-specific RBPs. Utilizing a CNN-based deep learning model with an ensemble of evolutionary features, RBProkCNN addresses the specificity of RBPs to individual species and lineage-specific families, providing more accurate predictions. Experimental validation of predicted RBPs, especially prokaryote-specific ones, will be crucial for confirming predictions and unravelling the functional significance of these proteins. Future endeavours should strive for continuous improvement, collaboration, and the integration of diverse data sources for a more comprehensive understanding of RBP functions in prokaryotes.

Funding information

This work was funded by ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi-110012, India (Grant Number-AGEDIASRISIL202101700188).

CRediT authorship contribution statement

Upendra Kumar Pradhan: Conceptualization, Data curation, Formal analysis, Methodology, Resources, Software, Writing – original draft, Writing – review & editing. **Sanchita Naha:** Data curation, Formal analysis, Resources, Software, Visualization, Writing – original draft. **Ritwika Das:** Data curation, Resources, Validation, Visualization. **Ajit Gupta:** Investigation, Resources, Software, Supervision, Writing – review & editing. **Rajender Parsad:** Investigation, Methodology, Project administration, Software, Supervision, Writing – review & editing. **Prabina Kumar Meher:** Conceptualization, Investigation, Methodology, Project administration, Software, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors sincerely acknowledge the Director, ICAR-IASRI, New Delhi for providing necessary computational facilities to carry out the research work. The authors also acknowledge the ASHOKA supercomputing facilities available at ICAR-IASRI, New Delhi.

Supplementary data

There is no supplementary data provided in the article.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.04.034.

References

[1] Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet 2014;15:829–45. https://doi.org/10.1038/nrg3813.
[2] Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 2008;582:1977–86. https://doi.org/10.1016/j.febslet.2008.03.004.
[3] Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. Nat Rev Mol Cell Biol 2018;19:327–41. https://doi.org/10.1038/nrm.2017.130.
[4] Hudson WH, Ortlund EA. The structure, function and evolution of proteins that bind DNA and RNA. Nat Rev Mol Cell Biol 2014;15:749–60. https://doi.org/10.1038/nrm3884.
[5] Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, et al. A large-scale binding and functional map of human RNA-binding proteins. Nature 2020;583:711–9. https://doi.org/10.1038/s41586-020-2077-3.
[6] Holmqvist E, Vogel J. RNA-binding proteins in bacteria. Nat Rev Microbiol 2018;16:601–15. https://doi.org/10.1038/s41579-018-0049-5.
[7] Mitchell SF, Jain S, She M, Parker R. Global analysis of yeast mRNPs. Nat Struct Mol Biol 2013;20:127–33. https://doi.org/10.1038/nsmb.2468.
[8] Oliveira C, Faoro H, Alves LR, Goldenberg S. RNA-binding proteins and their role in the regulation of gene expression in Trypanosoma cruzi and Saccharomyces cerevisiae. Genet Mol Biol 2017;40:22–30. https://doi.org/10.1590/1678-4685-GMB-2016-0258.
[9] Eisenreich W, Rudel T, Heesemann J, Goebel W. Link Between Antibiotic Persistence and Antibiotic Resistance in Bacterial Pathogens. Front Cell Infect Microbiol 2022;12:900848. https://doi.org/10.3389/fcimb.2022.900848.
[10] King AN, de Mets F, Brinsmade SR. Who's in control? Regulation of metabolism and pathogenesis in space and time. Curr Opin Microbiol 2020;55:88–96. https://doi.org/10.1016/j.mib.2020.05.009.
[11] Barquist L, Vogel J. Accelerating discovery and functional analysis of small RNAs with new technologies. Annu Rev Genet 2015;49:367–94. https://doi.org/10.1146/annurev-genet-112414-054804.
[12] Chakravarty S, Massé E. RNA-dependent regulation of virulence in pathogenic bacteria. Front Cell Infect Microbiol 2019;9.
[13] Holmqvist E, Wright PR, Li L, Bischler T, Barquist L, Reinhardt R, et al. Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. EMBO J 2016;35:991–1011. https://doi.org/10.15252/embj.201593360.
[14] Lazar V, Oprea E, Ditu L-M. Resistance, tolerance, virulence and bacterial pathogen fitness—current state and envisioned solutions for the near future. Pathogens 2023;12:746. https://doi.org/10.3390/pathogens12050746.
[15] Vestby LK, Grønseth T, Simm R, Nesse LL. Bacterial biofilm and its role in the pathogenesis of disease. Antibiot (Basel) 2020;9:59. https://doi.org/10.3390/antibiotics9020059.
[16] Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 2017;45:e84. https://doi.org/10.1093/nar/gkx059.
[17] Zheng J, Kundrotas PJ, Vakser IA, Liu S. Template-Based modeling of Protein-RNA interactions. PLoS Comput Biol 2016;12:e1005120. https://doi.org/10.1371/journal.pcbi.1005120.
[18] Yang Y, Zhan J, Zhao H, Zhou Y. A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid

binding prediction. Proteins 2012;80:2080–8. https://doi.org/10.1002/prot.24100.
[19] Yang Y, Zhao H, Wang J, Zhou Y. SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. Methods Mol Biol 2014;1137:119–30. https://doi.org/10.1007/978-1-4939-0366-5_9.
[20] Zhao H, Yang Y, Zhou Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. RNA Biol 2011;8:988–96. https://doi.org/10.4161/rna.8.6.17813.
[21] Sharan M, Förstner KU, Eulalio A, Vogel J. APRICOT: an integrated computational pipeline for the sequence-based identification and characterization of RNA-binding proteins. Nucleic Acids Res 2017;45:e96. https://doi.org/10.1093/nar/gkx137.
[22] Beckmann BM, Horos R, Fischer B, Castello A, Eichelbaum K, Alleaume A-M, et al. The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. Nat Commun 2015;6:10127. https://doi.org/10.1038/ncomms10127.
[23] Paz I, Kligun E, Bengad B, Mandel-Gutfreund Y. BindUP: a web server for non-homology-based prediction of DNA and RNA binding proteins. Nucleic Acids Res 2016;44:W568–74. https://doi.org/10.1093/nar/gkw454.
[24] Lam JH, Li Y, Zhu L, Umarov R, Jiang H, Héliou A, et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. Nat Commun 2019;10:4941. https://doi.org/10.1038/s41467-019-12920-0.
[25] Shazman S, Mandel-Gutfreund Y. Classifying RNA-binding proteins based on electrostatic properties. PLoS Comput Biol 2008;4:e1000146. https://doi.org/10.1371/journal.pcbi.1000146.
[26] Kumar M, Gromiha MM, Raghava GPS. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J Mol Recognit 2011;24:303–13. https://doi.org/10.1002/jmr.1061.
[27] Zhang X, Liu S. RBPPred: predicting RNA-binding proteins from sequence using SVM. Bioinformatics 2017;33:854–62. https://doi.org/10.1093/bioinformatics/btw730.
[28] Wang N, Zhang J, Liu B. iDRBP-EL: Identifying DNA- and RNA- Binding Proteins Based on Hierarchical Ensemble Learning. IEEE/ACM Trans Comput Biol Bioinform 2023;20:432–41. https://doi.org/10.1109/TCBB.2021.3136905.
[29] Wang N, Zhang J, Liu B. IDRBP-PPCT: identifying nucleic acid-binding proteins based on position-specific score matrix and position-specific frequency matrix cross transformation. IEEE/ACM Trans Comput Biol Bioinform 2022;19:2284–93. https://doi.org/10.1109/TCBB.2021.3069263.
[30] Feng J, Wang N, Zhang J, Liu B. iDRBP-ECHF: identifying DNA- and RNA-binding proteins based on extensible cubic hybrid framework. Comput Biol Med 2022;149:105940. https://doi.org/10.1016/j.compbiomed.2022.105940.
[31] Pradhan UK, Meher PK, Naha S, Pal S, Gupta S, Gupta A, et al. RBPLight: a computational tool for discovery of plant-specific RNA-binding proteins using light gradient boosting machine and ensemble of evolutionary features. Brief Funct Genom 2023;22:401–10. https://doi.org/10.1093/bfgp/elad016.
[32] Zheng J, Zhang X, Zhao X, Tong X, Hong X, Xie J, et al. Deep-RBPPred: predicting RNA binding proteins in the proteome scale based on deep learning. Sci Rep 2018;8:15264. https://doi.org/10.1038/s41598-018-33654-x.
[33] Zhang J, Chen Q, Liu B. iDRBP_MMC: Identifying DNA-Binding proteins and RNA-Binding proteins based on multi-label learning model and motif-based convolutional neural network. J Mol Biol 2020;432:5860–75. https://doi.org/10.1016/j.jmb.2020.09.008.
[34] Zhang J, Chen Q, Liu B. DeepDRBP-2L: a new genome annotation predictor for identifying DNA-Binding proteins and RNA-binding proteins using convolutional neural network and long short-term memory. IEEE/ACM Trans Comput Biol Bioinform 2021;18:1451–63. https://doi.org/10.1109/TCBB.2019.2952338.
[35] Zhang J, Yan K, Chen Q, Liu B. PreRBP-TL: prediction of species-specific RNA-binding proteins based on transfer learning. Bioinformatics 2022;38:2135–43. https://doi.org/10.1093/bioinformatics/btac106.
[36] Peng X, Wang X, Guo Y, Ge Z, Li F, Gao X, et al. RBP-TSTL is a two-stage transfer learning framework for genome-scale prediction of RNA-binding proteins. Brief Bioinforma 2022;23:bbac215. https://doi.org/10.1093/bib/bbac215.
[37] Yan K, Feng J, Huang J, Wu H. iDRPro-SC: identifying DNA-binding proteins and RNA-binding proteins based on subfunction classifiers. Brief Bioinforma 2023;24:bbad251. https://doi.org/10.1093/bib/bbad251.
[38] Nagarajan R, Gromiha MM. Prediction of RNA Binding Residues: an extensive analysis based on structure and function to select the best predictor. PLOS ONE 2014;9:e91140. https://doi.org/10.1371/journal.pone.0091140.
[39] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res 2023;51:D523–31. https://doi.org/10.1093/nar/gkac1052.
[40] Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26:680–2. https://doi.org/10.1093/bioinformatics/btq003.
[41] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402. https://doi.org/10.1093/nar/25.17.3389.
[42] Holm L, Sander C. Removing near-neighbour redundancy from large protein sequence collections. Bioinformatics 1998;14:423–9. https://doi.org/10.1093/bioinformatics/14.5.423.
[43] Vapnik V. Pattern recognition using generalized portrait method. Autom Remote Control 1963.
[44] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery,; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
[45] Breiman L. Random Forests. Mach Learn 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.
[46] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Light: A Highly Effic Gradient Boost Decis Tree 2017.
[47] Duchi JC, Hazan E, Singer Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. J Mach Learn Res 2011.
[48] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–232. https://doi.org/10.1214/aos/1013203451.
[49] Freund Y, Schapire R. A Short Introd Boost 1999.
[50] Breiman L. Bagging predictors. Mach Learn 1996;24:123–40. https://doi.org/10.1007/BF00058655.
[51] Kim Y. Convolutional Neural Networks for Sentence Classification. In: Moschitti A, Pang B, Daelemans W, editors. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1746–51. https://doi.org/10.3115/v1/D14-1181.
[52] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
[53] Brahma S. Improved Sentence Modeling using Suffix Bidirectional LSTM. arXiv: Learning 2018.
[54] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Moschitti A, Pang B, Daelemans W, editors. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics; 2014. p. 1724–34. https://doi.org/10.3115/v1/D14-1179.
[55] Chen R-C, Dewi C, Huang S-W, Caraka RE. Selecting critical features for data classification based on machine learning methods. J Big Data 2020;7:52. https://doi.org/10.1186/s40537-020-00327-4.
[56] Sandri M, Zuccolotto P. A bias correction algorithm for the gini variable importance measure in classification trees. J Comput Graph Stat 2008;17:611–28. https://doi.org/10.1198/106186008×344522.
[57] Sharma NK, Gupta S, Kumar A, Kumar P, Pradhan UK, Shankar R. RBPSpot: learning on appropriate contextual information for RBP binding sites discovery. iScience 2021;24:103381. https://doi.org/10.1016/j.isci.2021.103381.
[58] Sun X, Jin T, Chen C, Cui X, Ma Q, Yu B. RBPro-RF: use Chou's 5-steps rule to predict RNA-binding proteins via random forest with elastic net. Chemom Intell Lab Syst 2020;197:103919. https://doi.org/10.1016/j.chemolab.2019.103919.
[59] Mishra A, Khanal R, Kabir WU, Hoque T. AIRBP: accurate identification of RNA-binding proteins using machine learning techniques. Artif Intell Med 2021;113:102034. https://doi.org/10.1016/j.artmed.2021.102034.
[60] Wei Q, Zhang Q, Gao H, Song T, Salhi A, Yu B. DEEPStack-RBP: accurate identification of RNA-binding proteins based on autoencoder feature selection and deep stacking ensemble classifier. Knowl-Based Syst 2022;256:109875. https://doi.org/10.1016/j.knosys.2022.109875.
[61] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27:1226–38. https://doi.org/10.1109/TPAMI.2005.159.
[62] Pradhan UK, Meher PK, Naha S, Pal S, Gupta A, Parsad R. PlDBPred: a novel computational model for discovery of DNA binding proteins in plants. Brief Bioinforma 2023;24:bbac483. https://doi.org/10.1093/bib/bbac483.
[63] Motion GB, Howden AJM, Huitema E, Jones S. DNA-binding protein prediction using plant specific support vector machines: validation and application of a new genome annotation tool. Nucleic Acids Res 2015;43:e158. https://doi.org/10.1093/nar/gkv805.