A Comparative Study of Models for mRNA Subcellular Localization Prediction Using a Combined Dataset Benchmark *

Artyom Firstkov[1 [0000−0002−9636−9800]], Majid Forghani[1 [0000−0002−9443−3610]] and Michael Khachay[1 [0000−0003−3555−0080]]

1 N.N. Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences, 620108 Yekaterinburg, Russia

Abstract. Knowledge of the subcellular localization of mRNA is essential for understanding many cellular mechanisms, the disruptions of which are linked to severe diseases. Since wet-lab experiments for determining mRNA subcellular localization are time-consuming and labor-intensive, various machine learning models have been proposed to address these shortcomings. In this paper, we conducted a comparative study of models targeting prediction of mRNA subcellular localization in a cross-dataset benchmarking setting. We compiled the cross-benchmark dataset by fusing two well-known datasets, namely mRNALoc and RNAlight. Further, we introduced a stacking ensemble model, named mSP (mRNA stacking predictor), and confirmed its superior performance through computational experiments. To ensure a fair comparison, five state-of-the-art models were adopted from existing literature based on code availability and reproducibility of their training process. The results demonstrate that the proposed ensemble surpassed competing models across several performance measures, including overall accuracy, F1-score, and area under the curve.

Keywords: Machine learning · Multiclass classification · Subcellular localization · mRNA · Ensemble learning.

1 Introduction

The subcellular localization of messenger ribonucleic acid (mRNA) exerts a substantial influence on the regulation of gene expression [4], cell migration [20], and cellular differentiation [11]. The mislocalization of mRNA is linked with serious health concerns, including spinal muscular atrophy, Alzheimer's disease, Fragile X syndrome, embryonic disorders, and various types of cancer [19, 24].

Generally, the mRNA subcellular localization is assessed by wet-lab procedures such as Fluorescence in Situ Hybridization (FISH), its variants (e.g.,

* The work was performed as part of research conducted in the Ural Mathematical Center with the financial support of the Ministry of Science and Higher Education of the Russian Federation (Agreement number № 075-02-2026-737).

smFISH, seqFISH) [22] and Avidity-Based Proximity Labeling for RNA-Protein Interaction Profiling (APEX-RIP) [13]. Despite their reliable results and insights, the aforementioned wet-lab experiments are expensive, labor-intensive, and time-consuming. Consequently, numerous machine learning (ML) models have been developed to predict the localization of mRNA over the last decade [1,12,17,27]. A prominent advantage of such models is providing predictive insights into unseen and uncharacterized RNAs.

1.1 Related Works

Generally, modeling and prediction of mRNA subcellular localization can be presented in two ways: single-label multiclass classification [1] or multi-label localization task [25]. In a multiclass scenario, the goal is to assign the mRNA sequence to one of the subcellular location classes, whereas in a multilabel case, the sequence can be assigned to more than one class. In this paper, we address the single-label multiclass classification task.

To the best of our knowledge, Yan et al. [27] were the first to propose the mRNA subcellular localization as a classification problem. They built a deep learning (DL) model called RNATracker that leverages convolution, bidirectional LSTM, and attention layers. RNATracker attracted the attention of the scientific community to the problem of predicting RNA subcellular localization.

Zhang et al. [31] introduced iLoc-mRNA, a predictive model for mRNA localization of *Homo sapiens* based on Support Vector Machine (SVM) by employing RNALocate v1.0 [30] considering only those records that have exactly one subcellular localization. Their results indicated that iLoc-mRNA relatively outperformed RNATracker.

Garg et al. [12] implemented a model named mRNALoc by employing SVM and a one-versus-rest (OvR) strategy. The authors used RNALocate v2.0 [9] and built a dataset that served as a benchmark resource for several subsequent studies. To avoid overestimation of prediction capability and homology bias in prediction, they applied BLASTCLUST from the BLAST package [5] and generated a non-redundant mRNA dataset. The mRNA sequence was encoded into a numerical vector using pseudo-oligonucleotide composition or pseudo-k-tuple nucleotide composition (PseKNC) [7, 16] with k={2,3,4,5}. The authors compared the model with iLoc-mRNA across three classes (cytoplasm, endoplasmic reticulum, and nucleus), in the results of which their model achieved the highest overall true positive. It is worth mentioning that mRNALoc used sequences with low alignment identity during training while covering a wide range of mRNAs from all types of genes and isoforms, as well as all eukaryotes within RNALocate v2.0.

Yuan et al. [28] examined various classifiers, including SVM, logistic regression, and LightGBM, as well as DL models by leveraging the convolutional neural network (CNN), recurrent neural network (RNN), and a hybrid version by combining CNN and RNN. They employed k-mer frequency and one-hot encoding to embed the mRNA primary sequences. The results of computational experiments

confirmed that the model developed based on LightGBM achieved superior performance in comparison with the other two competing models, iLoc-mRNA and mRNALoc [12]. The authors named the proposed model RNAlight.

Musleh et al. [18] introduced MSLP (mRNA Subcellular Localization Predictor), an ML model based on CatBoost and OvR setup. The authors proposed an effective set of features by combining k-mer, pseudo-nucleotide composition, physicochemical properties, and 3D representation of the primary sequence obtained from Z-curve transformation. The computational experiments indicated that the proposed model outperformed mRNALoc across three out of five classes in terms of accuracy.

Wan et al. [26] introduced the application of DNABERT-2 [32] for advanced feature extraction in modeling subcellular localization, capturing more complex interrelationships that cannot be represented by handcrafted features. Besides the learned feature obtained from DNABERT-2, the authors also suggested the application of six handcrafted feature sets, including Z-curve [29], GC content, AT/GC ratio, GC skew, AT skew, and three-tuple nucleotide electron-ion interaction pseudopotential (EIIP) to ensure the multi-aspect representation of patterns in RNA sequences. They offered two primary classes by merging the mRNAs of different subcellular localization. By conducting a comprehensive computational experiment, they demonstrated that CatBoost with the fusion of both handcrafted and learned features surpassed other candidate models. The authors named the final model mRCat and benchmarked it against four state-of-the-art models: RNAlight, SubLocEP [15], RNATracker, and mRNALoc, exhibiting the superiority and stability of mRCat in terms of accuracy.

Aiming to enhance their previous model, MSLP, Musleh et al. [17] presented a new model called UMSLP (Unified mRNA Subcellular Localization Predictor). This model relies on several types of features, including k-mer, PseKNC (both with $k \in \{2, 3, 4, 5\}$), two representations of Z-curve, Pse-EIIP, dinucleotide physicochemical properties, and trinucleotide physicochemical properties. The authors adopted a series of dimensionality reduction techniques to avoid overfitting and eliminate collinear and redundant features. Their results revealed that Z-curve and k-mer play dominant roles in predicting subcellular localization. Their final published model was built based on CatBoost. They benchmarked UMSLP against mRNALocater [23], mRNALoc, and SubLocEP, demonstrating its robustness across multiple performance measures.

Although the aforementioned approaches appear to be promising and reach an acceptable degree of accuracy, to the best of our knowledge, each of them suffers from some shortcomings. In RNATracker, the mRNA sequence is embedded by one-hot encoding, which increases the input dimension size and represents the data in a sparse manner lacking the ordinal relationship and/or semantic meaning. As pointed out by the authors, the input variable length in the model training leads to a serious challenge that prevents the potential utilization of batch and batch normalization, likely due to incorporating the convolutional layer as the first layer. In the case of iLoc-mRNA, as stated by other researchers [12], merging nucleus, exosome, dendrite, and mitochondrion into one class seems to

be improper for a class definition, as they represent diverse subcellular localizations, each suggesting different patterns. mRCat relies on DNABERT-2, which is a pre-trained large language model. Prior training of such a model on the complete DNA dataset may have introduced data leakage by incorporating underlying features from the test set. mRNALoc did not provide a benchmark for the SVM model in comparison to other ML classifiers, such as random forest.

In addition to the aforementioned point, the authors of RNAlight mentioned that all DL models evaluated during their study achieved relatively poor performance in comparison with the ML models. They stated that this may have occurred due to the relatively small scale of the dataset for model development. Similarly, both MSLP and UMSLP are limited to modeling ML classifiers and handcrafted features. As stated by the authors, previous studies indicated that DL models did not attain a desired performance for predicting subcellular localization. However, DL already exhibited its high performance in many bioinformatics tasks that are designed to work with genetic sequences. An inadequate representation of RNA sequence via encoding/embedding, combined with a lack of suitable advanced architecture, likely limited the scope of previous conclusions. A proof of this statement has been recently proposed by LGLoc [1] that relies on SpliceBERT [6] and graph neural network [8].

This paper aims to conduct a comparative study of model performance in predicting mRNA subcellular localization using a cross-dataset benchmarking framework. Taking into account the availability of the code and reproducibility of the training process, we selected the following five models: mRNALoc [12], MSLP [18], UMSLP [17], RNAlight [28], mRCat [26]. Table 1 presents more details about the selected works.

Table 1. Characteristics of selected models regarding code availability and reproducibility of training process.

Article	Year	Data Source	Dataset	Model
mRNALoc [12]	2020	RNALocate v2.0	mRNALoc	SVM OvR
RNAlight [28]	2023	CeFra-seq [3]+APEX-Seq [10]	RNAlight	LightGBM
MSLP [18]	2023	RNALocate v2.0	mRNALoc, mLoc	CatBoost
mRCat [26]	2024	CeFra-seq+APEX-Seq	RNAlight	CatBoost
UMSLP [17]	2024	RNALocate v2.0	mRNALoc	CatBoost

Note: The "Data Source" column indicates the origin of the compiled dataset.

To the best of our knowledge and belief, while the current developed models for predicting mRNA subcellular localization have laid a solid foundation, there remains room for improvement of their performance. This motivated us to implement a novel model targeting the enhancement of prediction power.

1.2 Our Contribution

The contribution of this paper is as follows:

- A cross-data benchmark is compiled by merging and processing mRNALoc and RNAlight datasets. The dataset is further employed to conduct a fair and reproducible comparison of models.
- An evaluation of the selected candidate models is conducted against the benchmark dataset.
- A novel ensemble model, named mSP (mRNA stacking predictor), for predicting mRNA subcellular localization is proposed.

2 Problem Statement

The task can be formulated as mRNA primary sequence classification, which is indeed a supervised multiclass learning problem. Suppose the finite set $\Sigma = \{A, C, G, T\}$ denotes the nucleotide alphabet. Let Σ^* denote the set of all finite-length mRNA sequences. Note that Σ^* includes sequences of different lengths. Each input string $x \in \Sigma^*$ represents an mRNA primary sequence, i.e., a string of nucleotides from Σ with no additional structural annotation. Suppose $\mathcal{Y} = \{1, 2, \ldots, q\}$ denotes the finite set of class labels corresponding to distinct subcellular locations (organelles).

Given a labeled training dataset:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}, \quad \text{where} \quad x_i \in \Sigma^* \quad \text{and} \quad y_i \in \mathcal{Y}$$

the aim is to learn a predictive function:

$$f_\theta : \Sigma^* \longrightarrow \mathcal{Y}$$

where θ represents the function's parameters. f_θ maps an mRNA primary sequence into its corresponding class label.

The learning objective is to estimate vector parameter θ such that it minimizes the expected classification loss over the underlying data distribution $P(x, y)$. In empirical fashion, this is formulated as:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(f_\theta(x_i), y_i), \qquad (x_i, y_i) \in \mathcal{D},$$

where $\mathcal{L}(.,.)$ stands for a multiclass loss function, e.g., categorical cross-entropy.

3 Methodology

Figure 1 illustrates the schematic workflow of the proposed computational pipeline. Our investigation consists of two steps. In the first step, we compiled a novel benchmark dataset obtained by merging the mRNALoc and RNAlight datasets. Furthermore, a comparative analysis was conducted to evaluate the performance of five baseline models. In the second step, we developed an ensemble model using the one-versus-one (OvO) strategy and evaluated it alongside the results from the previous step. Each step is described in more detail in the following subsections.

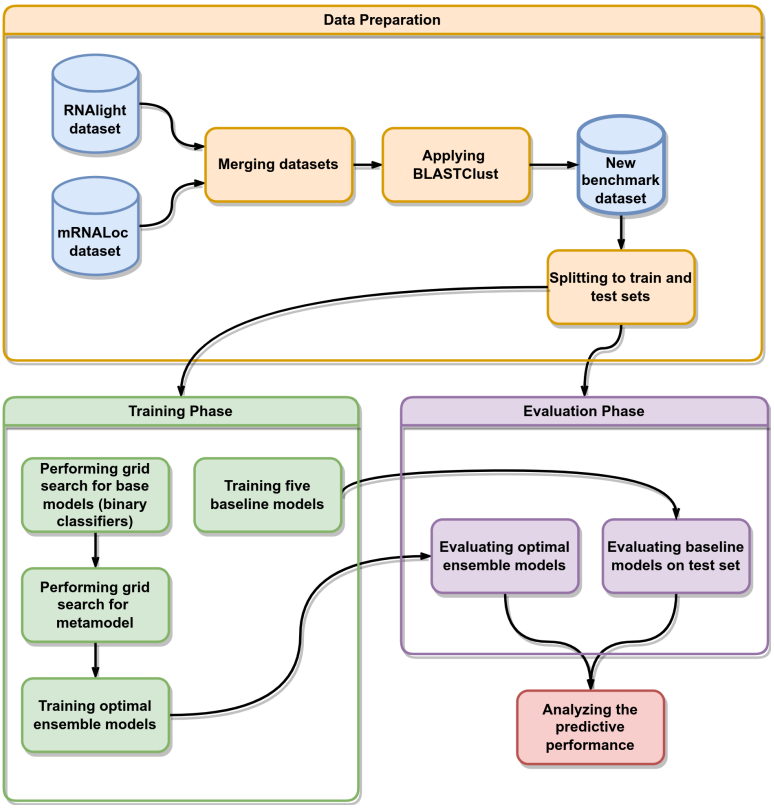

Fig. 1. Schematic workflow of the proposed computational pipeline. A novel benchmark dataset was compiled by merging the mRNALoc and RNAlight datasets. Further, the new dataset was employed to train and evaluate baseline models as well as develop an ensemble model for predicting mRNA subcellular localization.

3.1 Comparative study

Three of the five candidate models, selected for cross-dataset benchmarking, were trained on the mRNALoc dataset, while the remaining two models used the RNAlight dataset (see Table 1). To conduct a fair comparison and increase the dataset volume, we merged the mRNALoc and RNAlight datasets. The union of these datasets may lead to increasing the similarity between primary sequences, resulting in redundancy and data leakage. Therefore, we employed the NCBI BLASTClust program [5] with the recommended parameters [12] (-S 40 and -L 0.7) to keep sequences having alignment identity less than or equal to 40% over 70% or more of their full length. The longest sequence from each cluster was selected as the representative sample for constructing the newly compiled

benchmark dataset. The new dataset was randomly divided into training and test samples in a ratio of $4/5$ and $1/5$, respectively. Unlike many previous studies that ignore the distribution of sequence length in training and test samples, the data split was performed in such a way that the mean of sequence length remained consistent across both sets.

We trained five baseline models, including mRNALoc, MSLP, UMSLP, RNAlight, and mRCat, by adapting their source code regarding the new benchmark dataset. Training was conducted using the model configurations reported in the model's original paper. The training process of each model was reproduced as accurately as possible by the quality of the code, documentation, and supplementary materials. The performance was assessed using the following measures: overall accuracy (OA), macro F1-score (hereafter referred to as F1-score for brevity), precision (Pr), recall, Matthews correlation coefficient (MCC), area under the curve considering the OvO setting (OvO-AUC), and specificity (Sp). In this study, OA serves as the primary measure for performance evaluation, while supplementary measures play a supporting role, several of which account for class imbalance.

3.2 Model Development

It is a known fact that multiclass classification problems can face constraints that are often infeasible, where ensemble learning can overcome this limitation [14]. Recent studies indicated that ensemble models have potential to be employed in solving the mRNA subcellular localization problem [17, 28]. Typically, there are two strategies for implementing binary classifiers in an ensemble model for multiclass classification: one-versus-rest (OvR) and one-versus-one (OvO). In an OvR scenario, a binary classifier is trained for each class against other classes individually, whereas OvO aims at building a binary classifier for each pair of classes. A notable drawback of OvR is that binary classifiers are trained on highly imbalanced datasets, for each of which the negative class represents all objects belonging to the remaining $q - 1$ classes (where q is the number of classes). In the case of OvO, despite its computational complexity for constructing $q(q-1)/2$ classifiers, it relaxes the problem into several simpler, more focused binary subproblems, where the final solution is taken regarding ensemble decisions. Accordingly, OvO avoids increasing class imbalance when training binary classifiers, and it is more efficient to separate objects of two classes from each other than objects of one class from others. Therefore, we employed the OvO scenario for developing the proposed ensemble.

The diagram of our model is shown in Figure 2. Feature extraction critically impacts the performance of classifiers for mRNA localization classification. Thus, we selected a set of features relying on previous studies and our experience. Here, we employ k-mer, PseKNC [16] with $k \in \{2, 3, 4, 5\}$, Z-curve, and physicochemical properties of mRNA as suggested in MSLP [18] and UMSLP [17]. A feature vector of 6122 length was obtained by concatenating the aforementioned hand-crafted features. Seventy features were excluded due to insufficient variation across objects in the training dataset. The final 6052 features served as an

initial feature set for each of the binary classifiers within the proposed ensemble model. To avoid overfitting, we utilized principal component analysis (PCA) and joint mutual information maximization (JMIM) [2] to decrease the correlation, noise, and dimensionality in feature space.

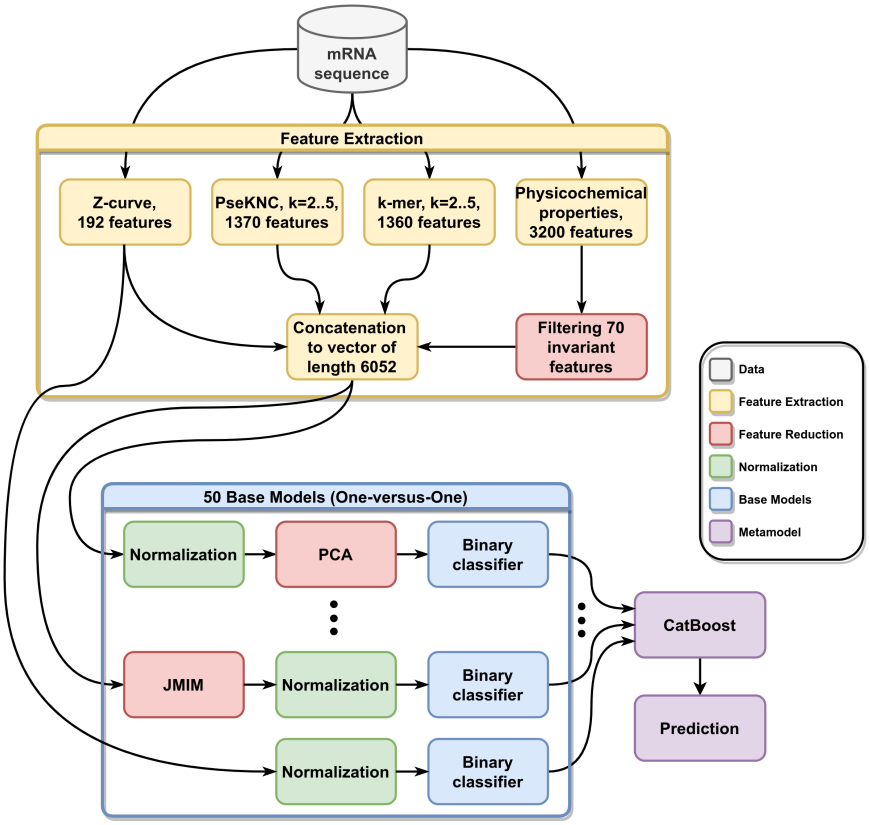

Fig. 2. Overall schema of the proposed model. It is worth mentioning that preprocessing (normalization and dimensionality reduction) was applied for each base model (binary classifier) individually. The proposed stacking ensemble uses class probability estimates to determine the label of an input instance.

Here, we considered two scenarios for the feature set configuration: the initial feature set subjected to dimensionality reduction; and the feature set represented by the Z-curve without any dimensionality reduction. It is noteworthy that the Z-curve was selected due to its effectiveness in discretizing the mitochondria class from other classes [17], as well as its low dimensionality. Further, we constructed

a dataset specified for each binary classifier, including its associated pair of labels. This dataset was normalized by removing the mean and dividing by the standard deviation. We applied a dimensionality reduction technique to the normalized dataset to generate the final dataset for training the associated classifier.

To optimize the ensemble's binary classifiers—the base models for each pair of classes—we used 5-fold stratified cross-validation and grid search for selecting the top models and their configurations. The grid search was conducted over the following seven classifiers: logistic regression (LR), k-nearest neighbors (KNN), support vector machine (SVM), light gradient boosting machine, random forest, adaptive boosting, and CatBoost. To assess the binary classifier performance, we calculated accuracy and balanced accuracy (BA). Here, BA served as the primary criterion for evaluating the performance during the grid search. Unlike accuracy, BA takes class imbalance into account and can be interpreted as the average recall of each class. Here, the final ensemble model was obtained by combining the top m-base models for each pair of classes according to the primary performance criterion, i.e., OA. Of note is that the classifier type and its configuration for the metamodel were also optimized via grid search.

4 Experiments & Results

Our experiment consists of two stages. The first stage is dedicated to identifying the optimal configurations for the proposed ensemble model, while at the second stage we benchmark it against competing models and demonstrate the robustness of our model.

As stated earlier, we merged mRNALoc and RNAlight datasets to compile a new benchmark dataset. The mRNALoc dataset contains 14909 samples of five classes, namely, cytoplasm (6376), nucleus (5831), endoplasmic reticulum (1426), extracellular region (855), and mitochondria (421). The RNAlight includes 5180 samples in two classes: cytoplasm (2924) and nucleus (2256). The two datasets comprised 20089 samples, while applying BLASTClust reduced this number to 19743 samples. Table 2 reports the class distribution of the new benchmark dataset. Unlike many previous studies that ignored the sequence length distribution, our benchmark dataset was divided to maintain similar mean sequence lengths across the training and test sets.

Table 2. The class distribution of the new benchmark dataset.

Set	Mean sequence length	Cytoplasm	Nucleus	Endoplasmic reticulum	Extracellular region	Mitochondria
Train	3454.5	7224	6449	1107	677	337
Test	3426.0	1807	1612	277	169	84

Table 3 represents the optimal binary classifier obtained in the results of grid search. It is worth mentioning that we considered two alternative cumula-

tive explained variance thresholds (0.95 and 0.99) for PCA to assess the impact of dimensionality reduction. The optimal setting for the cumulative explained variance threshold and resulting feature space dimension are presented in the "Reduction" and "Features" columns in Table 3, respectively. The experiment results indicated that, in some cases, only 1/50 of the initial feature set size is sufficient to achieve the highest balanced accuracy for binary classification. Majorly, SVM yielded the most robust results across all seven models. As expected and reported by previous studies, binary classification tasks involving the mitochondria class attained superior performance compared to the other class pairs. This is likely due to the origin of mitochondria [21].

Table 3. Optimal base models for ten pairs of classes derived via grid search.

Class Pair	Reduction	Features	Model	Train Score		Test Score	
				BA	Accuracy	BA	Accuracy
Cyt. - En. R.	PCA 0.99	733	SVM	0.7308	0.8701	0.7616	0.8887
Cyt. - Ext. R.	PCA 0.95	485	SVM	0.6341	0.8963	0.6309	0.8988
Cyt. - Mit.	PCA 0.99	731	LR	0.9840	0.9964	0.9864	0.9958
Cyt. - Nuc.	PCA 0.95	491	CatBoost	0.7791	0.7829	0.7873	0.7903
En. R. - Ext. R.	PCA 0.99	600	SVM	0.8517	0.8655	0.8538	0.8700
En. R. - Mit.	PCA 0.99	558	SVM	0.9864	0.9903	0.9845	0.9889
En. R. - Nuc.	PCA 0.95	432	SVM	0.8028	0.9087	0.8203	0.9153
Ext. R. - Mit.	PCA 0.99	523	SVM	0.9889	0.9921	0.9881	0.9921
Ext. R. - Nuc.	JMIM	11	KNN	0.6531	0.8555	0.6338	0.8501
Mit. - Nuc.	Z-curve	192	SVM	0.9806	0.9979	0.9881	0.9988

Cyt.: Cytoplasm, En. R.: Endoplasmic reticulum, Ext. R.: Extracellular region, Mit.: Mitochondria, Nuc.: Nucleus.

In our computational experiments, we explored two approaches in ensemble construction, including voting and stacking. Our initial metamodel for both approaches was assembled by combining the top-ten base models from Table 3. However, we implemented alternative ensemble variants in stacking by incorporating the top-m base models ($m \in \{1, 2, 3, 4, 5\}$) within each pair of classes and assessing the ensemble performance using OA. Table 4 lists the performance of voting and alternative variants of stacking ensembles obtained during our experiments. The results indicated that the stacking model (mSP)—based on the CatBoost classifier and comprising 50 base models—outperformed the others in terms of overall accuracy.

To demonstrate the predictive power of the ensemble model, we compared it with the following five models: mRNALoc, MSLP, UMSLP, RNAlight, and mRCat. As stated earlier, these models were selected due to the reproducibility criterion of the training process. Each model was trained from scratch and evaluated on train and test sets, respectively. Table 5 displays the results of computational experiments for evaluating the ensemble and competing baseline models on the new benchmark dataset. Despite both mRCat and RNAlight being

Table 4. Performance comparison of ensemble variants and baseline models via 5-fold cross-validation on train set.

Model	OA	F1-Score	Pr	Recall	MCC	OvO-AUC	Sp
mSP (CatBoost 50*)	**0.7433**	0.6764	0.7682	0.6437	**0.5720**	**0.9181**	**0.9108**
CatBoost 40*	0.7420	**0.6793**	0.7685	**0.6452**	0.5698	0.9168	0.9104
CatBoost 30*	0.7399	0.6736	0.7649	0.6403	0.5663	0.9141	0.9097
CatBoost 20*	0.7359	0.6585	0.7596	0.6271	0.5591	0.9092	0.9080
CatBoost 10*	0.7220	0.6344	0.7346	0.6095	0.5359	0.9008	0.9035
mRCat	0.7157	0.6200	0.7743	0.5806	0.5194	0.8928	0.9002
RNAlight	0.7122	0.5989	**0.8145**	0.5558	0.5123	0.8906	0.8978
Voting 10*	0.7111	0.5844	0.7566	0.5637	0.5104	0.8877	0.8980
UMSLP	0.6989	0.5709	0.7938	0.5420	0.4872	0.8818	0.8931
MSLP	0.6758	0.4561	0.7810	0.4364	0.4412	0.8582	0.8842
mRNALoc	0.6466	0.5071	0.7181	0.4945	0.3935	0.8382	0.8749

* The number following the meta-model name denotes the total number of base models.

originally designed for binary classification, notably, their multiclass adaptations outperformed the other baseline models. As shown in Table 5, the performance hierarchy of the models, developed on the mRNALoc dataset (i.e., UMSLP, MSLP, and mRNALoc), is consistent with the results reported in the UMSLP study. A similar situation is observed for the models developed using the mRNALoc dataset (i.e., mRCat and RNAlight), where the performance ranking in Table 5 aligns with the findings of the mRCat study.

In summary, the results indicated that our stacking model — mSP — outperformed the others in terms of all performance measures except precision. Given the pronounced class imbalance in the benchmark dataset, the observed improvement in the F1-score and MCC suggests that our model is more robust in this regard compared to the others.

Table 5. Comparative analysis of performance evaluation for proposed and baseline models on the test set.

Model	OA	F1-Score	Pr	Recall	MCC	OvO-AUC	Sp
mSP (this paper)	**0.7344**	**0.6744**	0.7547	**0.6480**	**0.5592**	**0.9208**	**0.9080**
mRCat	0.7298	0.6485	0.7775	0.6091	0.5456	0.9015	0.9054
RNAlight	0.7222	0.6145	0.7894	0.5728	0.5312	0.9033	0.9018
UMSLP	0.7133	0.5905	**0.8234**	0.5634	0.5133	0.8946	0.8981
MSLP	0.6880	0.4547	0.6640	0.4380	0.4625	0.8692	0.8887
mRNALoc	0.6404	0.5004	0.7757	0.4946	0.3828	0.8395	0.8726

5 Conclusion

In this paper, we performed a comparative study of state-of-the-art models for predicting mRNA subcellular localization across a combined dataset obtained

by merging mRNALoc and RNAlight datasets. By employing a one-versus-one strategy, we developed a stacking ensemble model on the new benchmark dataset. The model was constructed from 50 optimized binary classifiers targeting five subcellular compartments, namely, cytoplasm, nucleus, endoplasmic reticulum, extracellular region, and mitochondria. To optimize the training process, a distinct feature set was defined for each binary classifier using dimensionality reduction techniques, leading to significant reduction of the feature vector length. The results of computational experiments support that this optimization contributes to the enhancement of prediction quality.

To assess the performance of our ensemble, we benchmarked it against five well-known models from the literature, including mRNALoc, MSLP, UMSLP, RNAlight, and mRCat. The competing models were selected regarding the code availability and reproducibility of the training process. Unlike many other studies, before developing our own model, we adapted, trained from scratch, and evaluated the aforementioned baseline models. This ensures that all models are evaluated under identical training conditions. The results confirmed that the proposed ensemble model outperformed existing baseline models in terms of several performance measures, including the overall accuracy. Our study suggests that this robust performance enhancement can be attributed to a combination of several factors: application of the one-versus-one strategy, feature engineering, the customization of feature vectors for each binary classifier, the optimization of binary classifiers via a comprehensive grid search, and finally, the incorporation of a collection of top-m binary classifiers for each pair of classes in the ensemble construction. The developed model and new benchmark dataset are publicly available at github.com/Artyom438/mSP.

We plan to investigate the impact of other features, including both handcrafted and learned (by neural networks). Recent studies highlighted the efficiency of DL in modeling RNA subcellular localization. As part of our future work, we aim to explore advanced DL architectures and their impact on prediction accuracy.

Acknowledgments. This work was performed using the "Uran" supercomputer (Krasovskii Institute of Mathematics and Mechanics of the Ural Branch of the Russian Academy of Sciences).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akbari Rokn Abadi, S., Shahbakhsh, A., Koohi, S.: LGLoc as a new language model-driven graph neural network for mRNA localization. Scientific Reports **15**(1), 18709 (2025)
2. Bennasar, M., Hicks, Y., Setchi, R.: Feature selection using joint mutual information maximisation. Expert systems with applications **42**(22), 8520–8532 (2015)

3. Benoit Bouvrette, L.P., Cody, N.A., Bergalet, J., Lefebvre, F.A., Diot, C., Wang, X., Blanchette, M., Lécuyer, E.: CeFra-seq reveals broad asymmetric mrna and noncoding RNA distribution profiles in drosophila and human cells. Rna **24**(1), 98–113 (2018)
4. Besse, F., Ephrussi, A.: Translational control of localized mrnas: restricting protein synthesis in space and time. Nature reviews Molecular cell biology **9**(12), 971–980 (2008)
5. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L.: BLAST+: architecture and applications. BMC bioinformatics **10**(1), 421 (2009)
6. Chen, K., Zhou, Y., Ding, M., Wang, Y., Ren, Z., Yang, Y.: Self-supervised learning on millions of primary RNA sequences from 72 vertebrates improves sequence-based RNA splicing prediction. Briefings in bioinformatics **25**(3), bbae163 (2024)
7. Chen, W., Lin, H., Chou, K.C.: Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Molecular BioSystems **11**(10), 2620–2634 (2015)
8. Corso, G., Stark, H., Jegelka, S., Jaakkola, T., Barzilay, R.: Graph neural networks. Nature Reviews Methods Primers **4**(1), 17 (2024)
9. Cui, T., Dou, Y., Tan, P., Ni, Z., Liu, T., Wang, D., Huang, Y., Cai, K., Zhao, X., Xu, D., et al.: RNALocate v2.0: an updated resource for RNA subcellular localization with increased coverage and annotation. Nucleic acids research **50**(D1), D333–D339 (2022)
10. Fazal, F.M., Han, S., Parker, K.R., Kaewsapsak, P., Xu, J., Boettiger, A.N., Chang, H.Y., Ting, A.Y.: Atlas of subcellular rna localization revealed by APEX-Seq. Cell **178**(2), 473–490 (2019)
11. Gallicchio, L., Olivares, G.H., Berry, C.W., Fuller, M.T.: Regulation and function of alternative polyadenylation in development and differentiation. RNA biology **20**(1), 908–925 (2023)
12. Garg, A., Singhal, N., Kumar, R., Kumar, M.: mRNALoc: a novel machine-learning based in-silico tool to predict mRNA subcellular localization. Nucleic Acids Research **48**(W1), W239–W243 (2020)
13. Kaewsapsak, P., Shechner, D.M., Mallard, W., Rinn, J.L., Ting, A.Y.: Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. Elife **6**, e29224 (2017)
14. Khachay, M.: Committee polyhedral separability: complexity and polynomial approximation. Machine Learning **101**(1), 231–251 (2015)
15. Li, J., Zhang, L., He, S., Guo, F., Zou, Q.: SubLocEP: a novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning. Briefings in bioinformatics **22**(5), bbaa401 (2021)
16. Liu, B., Liu, F., Fang, L., Wang, X., Chou, K.C.: repDNA: a python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics **31**(8), 1307–1309 (2015)
17. Musleh, S., Arif, M., Alajez, N.M., Alam, T.: Unified mRNA subcellular localization predictor based on machine learning techniques. BMC genomics **25**(1), 151 (2024)
18. Musleh, S., Islam, M.T., Qureshi, R., Alajez, N.M., Alam, T.: MSLP: mRNA subcellular localization predictor based on machine learning techniques. BMC bioinformatics **24**(1), 109 (2023)

19. Otis, J.P., Mowry, K.L.: Hitting the mark: Localization of mRNA and biomolecular condensates in health and disease. Wiley Interdisciplinary Reviews: RNA **14**(6), e1807 (2023)
20. Park, Y., Page, N., Salamon, I., Li, D., Rasin, M.R.: Making sense of mRNA landscapes: translation control in neurodevelopment. Wiley Interdisciplinary Reviews: RNA **13**(1), e1674 (2022)
21. Roger, A.J., Muñoz-Gómez, S.A., Kamikawa, R.: The origin and diversification of mitochondria. Current Biology **27**(21), R1177–R1192 (2017)
22. Saw, P.E., Song, E.: RNA imaging technologies. In: RNA Therapeutics in Human Diseases, pp. 437–458. Springer (2025)
23. Tang, Q., Nie, F., Kang, J., Chen, W.: mRNALocater: enhance the prediction accuracy of eukaryotic mRNA subcellular localization by using model fusion strategy. Molecular Therapy **29**(8), 2617–2623 (2021)
24. Wang, E.T., Taliaferro, J.M., Lee, J.A., Sudhakaran, I.P., Rossoll, W., Gross, C., Moss, K.R., Bassell, G.J.: Dysregulation of mRNA localization and translation in genetic disease. Journal of Neuroscience **36**(45), 11418–11426 (2016)
25. Wang, X., Suo, W., Wang, R.: CSpredR: A multi-site mRNA subcellular localization prediction method based on fusion encoding and hybrid neural networks. Algorithms **18**(2), 67 (2025)
26. Wang, X., Yang, L., Wang, R.: mRCat: A novel CatBoost predictor for the binary classification of mRNA subcellular localization by fusing large language model representation and sequence features. Biomolecules **14**(7), 767 (2024)
27. Yan, Z., Lécuyer, E., Blanchette, M.: Prediction of mRNA subcellular localization using deep recurrent neural networks. Bioinformatics **35**(14), i333–i342 (2019)
28. Yuan, G.H., Wang, Y., Wang, G.Z., Yang, L.: RNAlight: a machine learning model to identify nucleotide features determining RNA subcellular localization. Briefings in bioinformatics **24**(1), bbac509 (2023)
29. Zhang, R., Zhang, C.T.: A brief review: The z-curve theory and its application in genome analysis. Current genomics **15**(2), 78–94 (2014)
30. Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., Yang, H., Hu, Z., Zhang, L., Hu, C., et al.: RNALocate: a resource for RNA subcellular localizations. Nucleic acids research **45**(D1), D135–D138 (2017)
31. Zhang, Z.Y., Yang, Y.H., Ding, H., Wang, D., Chen, W., Lin, H.: Design powerful predictor for mRNA subcellular location prediction in homo sapiens. Briefings in Bioinformatics **22**(1), 526–535 (2021)
32. Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., Liu, H.: DNABERT-2: Efficient foundation model and benchmark for multi-species genome. arXiv preprint arXiv:2306.15006 (2023)