Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31432

Construction of a Machine Learning Pipeline Based on NMR Data: Analysis and Determination of the Primary Structure of Single-Stranded RNA

Shi Chen

Sunway College, Monash University Foundation Year, Petaling Jaya, Malaysia
Alisa6662025@outlook.com

Abstract. This study aims to establish a machine learning pipeline for determining and analyzing the primary structure using NMR data of single-stranded RNA. The pipeline consists of two steps, stage 1 uses an RNA binary classification model, and stage 2 uses an A/U/G/C four-classification recognition and sorting model. During the experiment, the single-stranded RNA NMR data collected from BMRB and NP-MRD data sources were processed using category imbalance calculation and SMOTE oversampling methods. Models such as random forest and gradient boosting, 5-fold cross-validation, Wilson score confidence interval, and generalization ability evaluation were used to determine the generalization ability of the models and whether overfitting occurred. Results show that the best model for stage1 is the Random Forest model (with 30 features)with an accuracy rate of 90.48%; the best model for stage2 is the Gradient Boosting model (100 trees, depth 5, learning rate 0.1) with an accuracy rate of 96.30% on the independent test set. And in the feature engineering of the stage2 model, four H6-H5 difference features were added, which cut down the confusion between C and U and improved the accuracy of the model. This machine learning pipeline can predict RNA sequences of 8-20 nucleotides based on NMR data.

Keywords: Machine learning, Single-stranded RNA recognition, NMR data, Primary structure determination.

1. Introduction

RNA is vital for life sciences. Beyond its core functions in gene expression and protein synthesis, it has gained new applications in recent years, such as RNA interference technology and mRNA vaccines. However, the biological functions of RNA are closely tied to its precise three-dimensional structure, making the characterization of RNA structure key to understanding its functions. Traditional structural analysis methods such as X-ray crystallography and cryo-electron microscopy can provide high-resolution data, but they face limitations when analyzing smaller RNA molecules or specific RNA secondary structures.

Nuclear Magnetic Resonance (NMR) technology has unique advantages: it provides insights into RNA structures in solution without crystallization and captures the dynamic behavior of molecules.

However, NMR spectrum analysis is complex and manual interpretation is time-consuming, new automated methods are needed to speed up research.

In recent years, machine learning, as a powerful data analysis tool, has shown great potential in NMR data interpretation.

Machine learning algorithms such as DNN and SVM work well in NMR spectrum peak-picking [1], and their application in NMR chemical shift prediction has boosted biomolecular structural analysis efficiency [2]. In RNA research, machine learning has driven progress in secondary structure prediction and drug development [3], while studies combining RNA classification with NMR data have grown, such as RNA NMR chemical shift data used to characterize excited states [4]. Machine learning also excels at handling complex NMR data like RNA-solvent interference, offering new ideas for RNA classification and identification [5].

This study explores an automatic RNA identification method based on NMR chemical shift data and machine learning. It seeks to build an efficient model for distinguishing RNA from non-RNA molecules, use chemical shift features in NMR spectra, and combine machine learning algorithms for rapid classification and sequence identification of single-stranded RNA.

2. Research methods

2.1. Data collection

This study retrieves data from the Biological Magnetic Resonance Data Bank (BMRB) database [6] and The Natural Products Magnetic Resonance Database (NP-MRD) database [7].

In the phase one RNA binary classification dataset, there are RNA data of 171 sequences from BMRB and 352 non-RNA organic molecules from NP-MRD, totaling 523 compounds. The training set contains 418 compounds, the test set contains 105 compounds, and the class balance is 1:2.1. The class imbalance ratio, calculated based on [8], is approximately 2.06 based on the aforementioned sample counts.

The dataset for base sequence prediction in stage 2 includes 105 RNA sequences from BMRB with 1314 residues, following the inclusion criteria of having complete ^{1}H and ^{13}C chemical shifts, containing standard nucleotides (A,U,G,C), and having a sequence length of 8-20 nucleotides. The base distribution includes G 433 33.7%, C 356 27.1%, A 288 21.9%, U 227 17.3%. These 105 data points are split into 80 training, 10 validation, and 15 independent test data points totaling 189 residues.

2.2. Two-stage machine learning pipeline

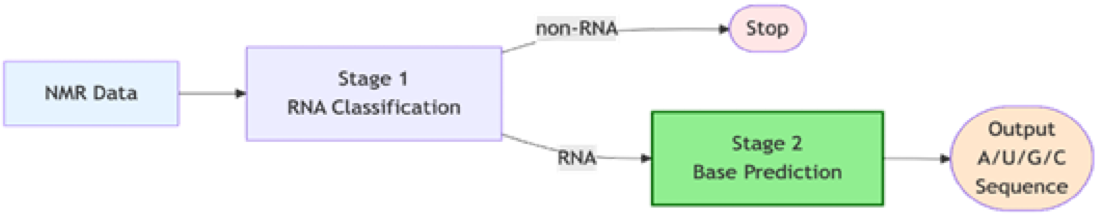

Figure 1. Flowchart of the two-stage machine learning pipeline (picture credit: original)

Figure 1 presents the flow chart of the study's two-stage machine learning pipeline. Stage 1 involves RNA binary classification to distinguish RNA from non-RNA compounds. Fifteen initial baseline features were optimized to 30 final features, including 5 basic statistics (mean, standard deviation, minimum, maximum and median), 7 regional peak densities across 0-3, 3-6, 6-8 and 8-10 ppm, 5 RNA-specific features such as aromatic ratio and sugar ratio, 3 peak spacing features (mean, standard deviation and maximum) and 10 additional statistics. A random forest model with 200 trees and a depth of 15 was used. Regional peak density is defined as the number of peaks falling within a specific chemical shift interval divided by the total number of peaks, where the intervals are numbered 1 to 4 corresponding to the aforementioned chemical shift ranges.

Stage 2 is one-dimensional base sequence prediction, with the goal of predicting the base type (A/U/G/C) of each residue. A key challenge encountered in the experiment is the confusion between C and U because both are pyrimidines. Therefore, in the experiment, the features evolved from the original 40 features (residue-level chemical shifts) to 52 features with the addition of the H6-H5 difference feature. The core innovation is the H6-H5 chemical shift difference. Due to the difference between C and U at the C4 position, one is -NH_2 and the other is =O, it was found that the H6-H5 of cytidylate is approximately 2.4 ppm, and the H6-H5 of U is approximately 2.0 ppm. Therefore, four H6-H5 difference features were extracted.The formula for the absolute difference is: $\Delta_{abs} = |\delta(H6) - \delta(H5)|$. The normalized difference is defined via the formula: $\Delta_{norm} = \frac{\delta(H6)-\delta(H5)}{\delta(H6)}$. The chemical shift ratio is expressed as: $R_{H6/H5} = \frac{\delta(H6)}{\delta(H5)}$. For the binary indicator, its definition is given as:

$$I_{H6,H5} = \begin{cases} 1, \ if \ \delta(H6) \ and \ \delta(H5) \ both \ exist \\ 0, \ otherwise \end{cases} \tag{1}$$

Among them, $\delta(H6)$ and $\delta(H5)$ are the chemical shifts of the H6 and H5 protons, respectively (unit: ppm).

This nucleotide chemical shift feature engineering strategy follows recent feature optimization practices, and fine feature engineering is effective for small dataset predictive performance, supporting the RNA chemical shift feature design.

The 52 features comprise 8 direct chemical shift features, 4 H6-H5 difference-related features, 24 sequence-level statistics, 3 purine/pyrimidine discriminators, and 13 additional derived features. The optimal model is gradient boosting with optimized key hyperparameters, which adopts SMOTE oversampling [9] to address class imbalance. SMOTE is effective for handling such imbalance in RNA-related multi-omics data [10].

2.3. Model training and validation

This study’s training protocol uses a standard CPU-only workstation (no GPU acceleration needed), with a software environment including Python 3.13, scikit-learn 1.7.2, and imbalanced-learn. Feature standardization used StandardScaler (fitted only on the training set), and 5-fold stratified cross-validation was used. Hyperparameters were set based on literature recommendations and empirical tuning, with the model taking about 5 minutes for stage 1 training and 15 minutes for stage 2. Both k-fold cross-validation accuracy and macro-average F1 score were calculated using references [11,12].

A three-layer validation protocol was used, including cross-validation, a validation set, and an independent test set. Cross-validation supports model selection, performance estimation, and algorithm comparison to find the optimal model. The validation set (10 RNA sequences) is exclusive to stage 2, and only confirms the final model configuration after cross-validation. The independent test set, separated and preserved prior to any model training, provides a real-world evaluation of final performance. This strategy follows best practices in molecular prediction [13], ensuring strict separation of dataset roles, while the independent test set reduces overly optimistic confidence interval bias from cross-validation data correlation [14].

To prevent data leakage, this study used rigorous measures to ensure the machine learning model's reliability and generalizability for RNA primary structure analysis. An independent test set was split from the original dataset using stratified random sampling (random_state=2025) prior to model training, to maintain consistent class distribution. StandardScaler was only fitted on the training set, and its transformation parameters were applied to the test set to avoid using unseen data information. SMOTE oversampling was limited to cross-validation training folds and never used for validation or test data. Programmatic verification confirmed 0% sequence overlap between the independent test set and training set, while base composition similarity between the two subsets was measured using cosine similarity—a threshold of less than 95% ensures sufficient base distribution difference to prevent indirect data leakage.

To verify the model's generalization ability, the accuracy difference $\Delta\mathrm{Acc}$ (defined as test accuracy minus cross-validation accuracy) and 95% confidence interval overlap were used as evaluation criteria— $\Delta\mathrm{Acc} \geq 0$ and overlapping intervals indicate good generalization and no overfitting. In Stage 2, $\Delta\mathrm{Acc}$ was 1.83% (96.30% minus 94.47%), falling within normal statistical variation. For reliable confidence interval calculation and statistical reliability of accuracy estimates, we adopted the Wilson method [15], which outperforms normal approximation for small-sample interval estimation and is recommended for machine learning cross-validation accuracy confidence intervals [16]. Statistical consistency between cross-validation and independent test set performance was verified via 95% confidence interval overlap; overlapping intervals confirm performance differences are within normal statistical variation, supporting the model's full generalization without overfitting.

3. Results and discussion

3.1. Stage 1 performance: RNA binary classification model

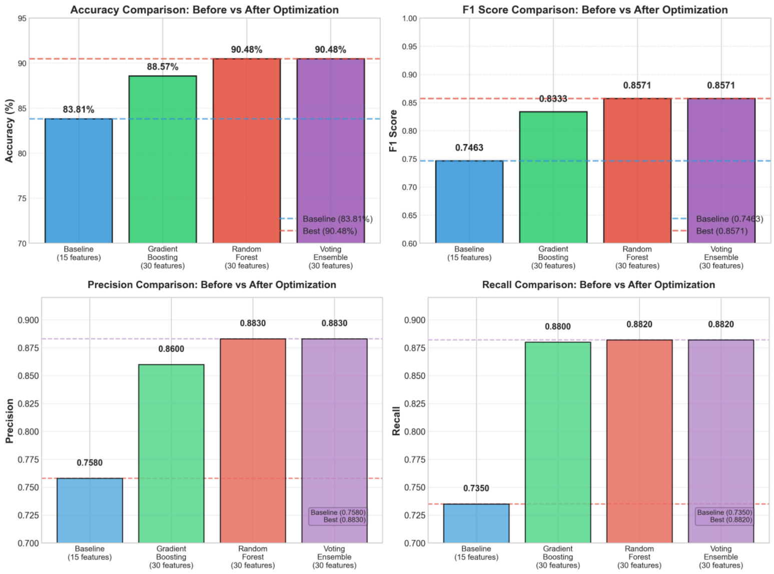

Figure 2. RNA recognition effect of the binary classification model (picture credit: original)

Based on the experimental results in Figure 2, feature optimization increased from 15 baseline dimensions to 30 final dimensions. The Random Forest (30 features) and Voting Ensemble (30 features) models achieved the optimal performance with an F1 score of 0.857. The accuracy of the Random Forest (30 features) and Voting Ensemble (30 features) was 6.67%higher than that of the baseline model (GB,15 features). The Random Forest was selected because of its superior performance(6.67% higher than the baseline), high computational efficiency, and, under the condition of having the same performance as the Voting Ensemble, the single model is more concise, easier to deploy and interpret.

Compared with the baseline features, the optimized features perform significantly better—this improvement comes from regional peak density features (+2-3%), RNA-specific features (aromatic/sugar ratio) (+2-3%), peak spacing features (+1-2%), and model optimization (switching GB to RF) (+1%).

Five-fold stratified cross-validation gave consistent performance, with an average accuracy of 89.52%±1.24% and a fold range of 88.1%-91.2%. The low standard deviation means the model is stable and robust.

For the confusion matrix analysis, on the independent test set (about 105 samples), the true negatives (correctly identified non-RNAs) account for 91.2%, the true negatives (correctly identified non-RNAs) make up 91.2%, true positives (correctly identified RNAs) make up 89.7%, false positives (non-RNAs misclassified as RNAs) make up 8.8%, and false negatives (RNAs misidentified as non-RNAs) make up 10.3%. The false negatives are mostly short RNA sequences lacking sufficient chemical shift information (fewer than 10 nucleotides).

3.2. Stage 2 performance: base sequence prediction

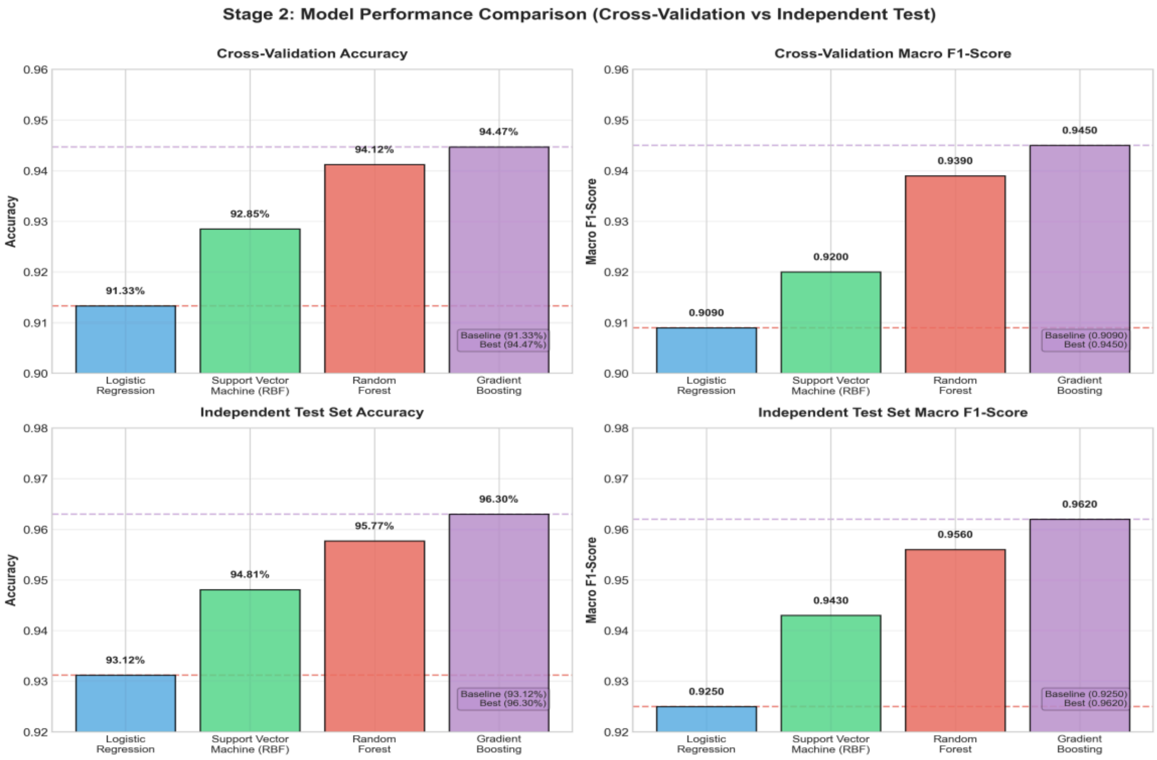

Figure 3. Comparison of RNA sequence predictions by four models (picture credit: original)

In Figure 3, gradient boosting with version 2 features (including H6-H5 difference) (100 trees, depth 5, learning rate 0.1) achieved the best performance in both cross-validation (94.47%) and the independent test set (96.30%). Gradient boosting (especially XGBoost) has been proven to outperform random forests and deep neural networks in molecular property prediction tasks, with higher computational efficiency [17], which meets the requirements of this study for high performance and interpretability.

Table 1. comparison of feature versions

Version	Number of features	Cross-validation	Test accuracy	Proportion of C/U errors
V1	40	93.00%	93.50%	60%
V2	52	94.47%	96.30%	29%

The H6-H5 feature effectively reduces C/U nucleotide confusion by more than 50% while increasing the cross-validation accuracy by 1.5% and the test accuracy by 2.8%, as shown in Table 1.

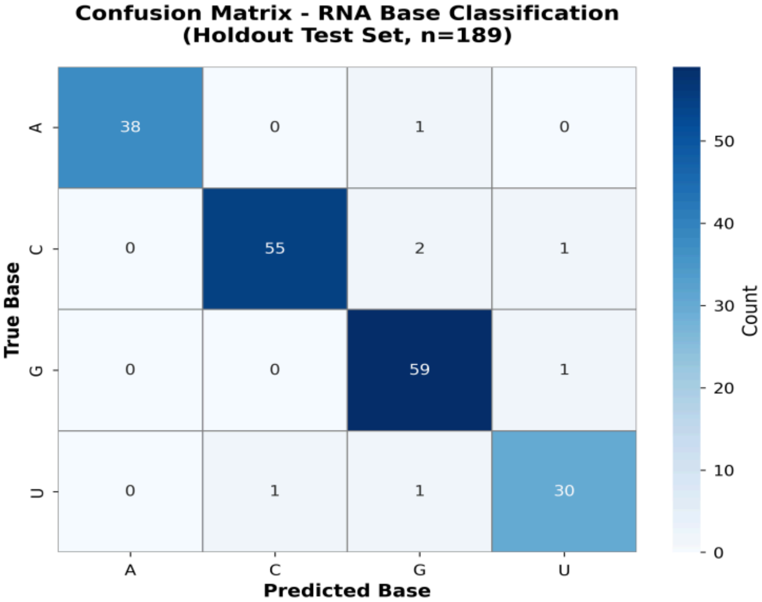

Figure 4. Confusion matrix of stage 2 base sequence prediction on the independent test set (picture credit: original)

Analysis of 7 misclassification is in figure 4 that there were 2 cases of confusion between C and U (28.6%), due to the similarity of pyrimidine structures and the boundary H6-H5 values. There was 1 case of confusion between A and G (14.3%), attributed to similar H8 chemical shifts and purine confusion. C was mistakenly identified as G in 2 cases (28.6%), with the error being atypical chemical shifts. The confusion between U and G occurred a total of 2 times (28.6%), which were respectively caused by the rare error of uracil being mistaken for guanine and atypical displacement. A key finding is that there were no other confusion types such as G being mistaken for A or A being mistaken for C, which indicates that the model can clearly distinguish between most base pairs. The main challenges lie in the boundary between C and G and a small number of cross-class errors. For the analysis of the confusion matrix in multi-class classification, this study followed the standard method for calculating TP, TN, FP, and FN for each category [18], and comprehensively evaluated the model performance using the macro-averaged F1 score. This method offers a more comprehensive evaluation than a single accuracy rate for imbalanced multi-classification problems [19].

For how well the H6-H5 feature works, the v1 model without it has about 9 C/U errors (out of roughly 15 total errors). The v2 model with this feature has 2 C/U errors among 7 total errors—cutting C/U errors by over 70% compared to the v1 model.

The gradient boosting model gives probability estimates for each prediction. On the independent test set, the average confidence level across all predictions is 98.92%—99.31% for correct ones and 68.45% for incorrect ones. For the confidence threshold, all 7 incorrect predictions have confidence levels below 75%. This means the model's confidence level can act as a reliability indicator to flag uncertain predictions for manual review.

Among the 15 independent test sequences, 11 (73.3%) achieved 100% accuracy (all residues correct), 3 had one error each (92.2% accuracy), and 1 had two errors (85.7% accuracy). The high proportion of perfect predictions shows practical value for fully automated RNA base identification.

3.3. Statistical verification

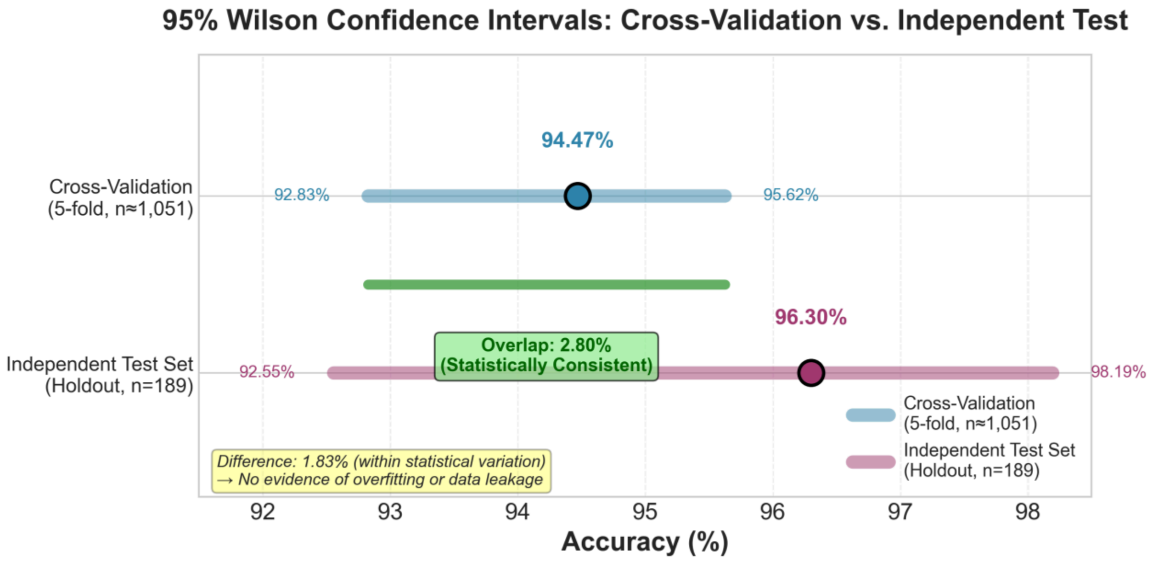

Figure 5. Verification of the generalization ability of the stage two model: Comparison of confidence intervals between cross-validation and independent test sets (picture credit: original)

The 95% confidence interval is shown in Figure 5, using the Wilson method. Cross-validation: 94.47% [92.83%,95.62%] (n≈1,051). Independent testing: 96.30% [92.55%,98.19%] (n=189). The difference is 1.83%, which is within the normal statistical variation. The conclusion is that there is no evidence of overfitting or data leakage.

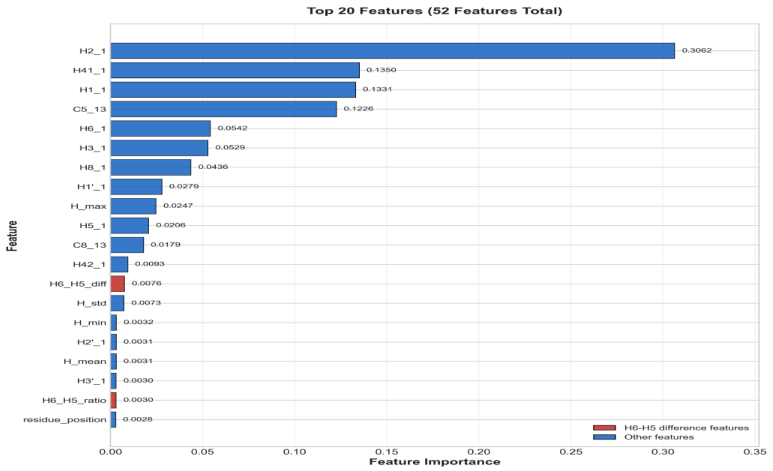

Figure 6. Ranking of the top 20 feature importances among 52 features (picture credit: original)

The H6-H5 difference features rank 13th (H6_H5_diff) and 19th (H6_H5_ratio) in the gradient boosting model which are shown in Figure 6. Although their overall rankings are not high, the ablation study shows that after removing these features, the C/U confusion increases significantly, and the accuracy rate drops from 96.30% to 93.36% (-2.94%, $p<0.001$), which proves that they play a key role in distinguishing pyrimidine bases.

4. Conclusion

Methodologically, the characteristic of the H6-H5 chemical shift difference effectively distinguishes cytosine from uracil and reduces C/U confusion by more than 50%, which represents an innovation in feature engineering and provides a reference idea for machine learning analysis of NMR data; the two-stage pipeline cascaded classifier architecture which first identifies RNA and then predicts bases can be applied to hierarchical classification tasks of other biomolecules.

Achieving an accuracy of 96.30% on a dataset with only 105 sequences and 1314 residues demonstrates that classical machine learning methods still have competitiveness in structured NMR data, showing high performance with small samples.

In practical applications, automated base recognition can reduce the time for manual identification from several hours to several minutes, and is particularly suitable for the rapid analysis of standard RNA sequences. High-confidence predictions can be used to mark abnormal chemical shifts or potential identification errors, assisting in experimental verification.

This pipeline has limitations. Due to the scarcity of complete RNA data containing both NMR and sequencing data in the BMRB database, the sequence length that can be analyzed by this pipeline model is limited to 8-20 short chains of standard nucleotides (A,U,G,C), excluding modified bases. It mainly operates under NMR conditions of D_2O and 25°C, requires complete ^1H and ^{13}C chemical shift data, and its robustness against chemical shift outliers (such as metal binding and extreme pH) has not been fully verified.

In the future, it will be able to expand to modified nucleotides (m^6A, ψ, etc.) and integrate NOESY/TOCSY multidimensional NMR data, incorporate ^2D NMR correlation peak information, and test the generalization ability for longer RNA sequences (>20 nucleotides). A recent systematic review on RNA structure prediction points out that the mixture-of-experts approach integrating deep learning and physical models can improve out-of-distribution generalization ability, which provides a direction for the model improvement in this study. In addition, developing online prediction tools and transferring learning to DNA and proteins are also important research directions.

References

[1] Li, D.W., Hansen, A. L., Bruschweiler-Li, L., et al (2022). Fundamental and practical aspects of machine learning for the peak picking of biomolecular NMR spectra. Journal of Biomolecular NMR, 76(4), 649–657. https: //doi.org/10.1007/s10858-022-00393-1

[2] Cortés, I., Cuadrado, C., Hernández Daranas, A., et al (2023). Machine learning in computational NMR-aided structural elucidation. Frontiers in Natural Products, 2, 122426. https: //doi.org/10.3389/fnphe.2023.1122426

[3] Sato, K., & Hamada, M. (2023). Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. Briefings in Bioinformatics, 24(4), 1–13. https: //doi.org/10.1093/bib/bbad186

[4] Wang, Y., Han, G., Jiang, X., et al. (2021). Chemical shift prediction of RNA imino groups: application toward characterizing RNA excited states. Nature Communications, 12, 1595. https: //doi.org/10.1038/s41467-021-21840-x

[5] Kuhn, S. (2022). Applications of machine learning and artificial intelligence in NMR. Magnetic Resonance in Chemistry, 60(12), n/a. https: //doi.org/10.1002/mrc.5310

[6] Hoch, J. C., Baskaran, K., Burr, H., et al (2023). Biological magnetic resonance data bank. Nucleic Acids Research, 51(D1), D368–D376. https: //doi.org/10.1093/nar/gkac1050

[7] Wishart, D. S., Sajed, T., Pin, M., et al (2025). The Natural Products Magnetic Resonance Database (NP-MRD) for 2025. Nucleic Acids Research, 53(D1), D700–D708. https: //doi.org/10.1093/nar/gkae10

[8] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https: //doi.org/10.1109/TKDE.2008.239

[9] Chawla, N. V., Bowyer, K. W., Hall, L. O., et al (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https: //doi.org/10.1613/jair.953

[10] Yang, Y., & Mirzaei, G. (2024). Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification. PloS one, 19(2), e0293607. https: //doi.org/10.1371/journal.pone.0293607

[11] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the International Joint Conference on Artificial Intelligence (pp. 1137-1145). https: //www.ijcai.org/Proceedings/95-2/Papers/016.pdf

[12] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https: //doi.org/10.1016/j.ipm.2009.03.002

[13] Saha, U. S., Vendruscolo, M., Carpenter, A. E., et al (2024). Step Forward Cross Validation for Bioactivity Prediction: Out of Distribution Validation in Drug Discovery. bioRxiv: the preprint server for biology, 2024.07.02.601740. https: //doi.org/10.1101/2024.07.02.601740

[14] Robinson, M. C., Glen, R. C., & Lee, A. A. (2020). Validating the validation: reanalyzing a large-scale comparison of deep learning and machine learning models for bioactivity prediction. Journal of computer-aided molecular design, 34(7), 717–730. https: //doi.org/10.1007/s10822-019-00274-0

[15] Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. Journal of the American Statistical Association, 22(158), 209–212. https: //doi.org/10.1080/01621459.1927.10502953

[16] Bayle, P., Bayle, A., Janson, L., et al (2020). Cross-validation confidence intervals for test error. In Advances in Neural Information Processing Systems (Vol. 33). https: //proceedings.neurips.cc/paper/2020/file /bce9abf229ffd7e570818476ee5d7dde-Paper.pdf

[17] Boldini, D., Grisoni, F., Kuhn, D., et al (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. Journal of Cheminformatics, 15, 73. https: //doi.org/10.1186/s13321-023-00743-7

[18] Markoulidakis, I., Rallis, I., Georgoulas, I., et al (2021). Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. Technologies, 9(4), 81. https: //doi.org/10.3390/technologies9040081

[19] Bharathi. (2021). Latest guide on confusion matrix for multi-class classification. Analytics Vidhya. https: //www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/