IT1244 Project Report: mRNA Classification

National University of Singapore

Team 20

Bagadia Litisha Vivekananda, Bui Nguyen Bao Khanh, Law Cheuk Yin, Tran Viet Khoa

1- Introduction

Messenger RNA (mRNA) is a special type of RNA that carries the genetic code from DNA to ribosomes, where proteins are made. It plays a crucial role in modern biology and has had many applications, most notably in COVID-19 vaccines, which use synthetic mRNA to instruct the body's cells to make specific vital protein (National Human Genome Research Institute, 2025).

Consequently, the classification task of mRNA has become increasingly important. Yet, traditional laboratory methods for mRNA separation, such as magnetic capture technologies, remain costly, time-consuming, and prone to various challenges (QIAGEN, 2025; Thermo Fisher Scientific, 2025). Applying Machine Learning (ML) could make this process faster, cheaper, and more efficient.

Several models are developed for mRNA classification. However, this task still remains challenging, where:

- The hybrid CNN–LSTM model proposed by Tasdelen & Sen (2021) has a small dataset and low specificity.
- The EDCLoc model by Deng, Jia, and Yi (2024) faces class imbalance issues and is susceptible to overfitting.
- The CNN-based model by Wen et al. (2019), though highly accurate, relies mainly on k-mer frequency features, which lose positional and long-range sequence information.
- The MSLP model by Musleh et al. (2023) uses engineered k-mer and physicochemical features but remains limited in capturing global contextual patterns.

The primary aim of this project is to develop a ML model capable of accurately distinguishing mRNA sequences from non-mRNA sequences using nucleotide data. By evaluating multiple models and feature extraction techniques, we aim to identify an efficient, interpretable, and generalizable framework for mRNA classification that can reduce dependence on traditional, resource-intensive laboratory methods.

2 - Dataset

The training dataset consists of DNA sequences paired with binary class labels, where 1 indicates mRNA and 0 indicates non-mRNA. A total of 14 286 labelled sequences were included: 9,244 non-mRNA (64.62%) and 5,062 mRNA (35.38%).

This uneven distribution shows a 1.8 : 1 ratio between non-mRNA and mRNA sequences. Such imbalance biases classifiers toward the majority class, often producing deceptively high accuracy while lowering recall or sensitivity for true mRNA instances.

To address this, all chosen models were trained with class_weight = "balanced_subsample".

This parameter automatically re-weights each class inversely to its frequency in every bootstrap subsample. Hence, each class contributes proportionally to the impurity or loss computation, reducing bias during training.

In addition, evaluation metrics such as Matthews Correlation Coefficient (MCC) and ROC-AUC were implemented, ensuring fairer performance assessment across imbalanced data.

Moreover, approximately 2 667 (18.7%) sequences contain non-canonical IUPAC symbols. These letters represent uncertain bases introduced during sequencing, which complicate k-mer extraction and downstream feature engineering.

Unresolved ambiguities introduce noise and obscure biological signals, particularly in mRNA regions with codon-specific triplet patterns.

We resolved this via equal-probability substitution. Each ambiguous base was replaced by one valid nucleotide chosen uniformly at random, with a fixed per-sequence random seed to ensure reproducibility. This approach effectively transforms ambiguous bases into biologically plausible ones without biasing substitutions.

3 - Feature Extraction

Length Distribution

The sequence length feature captures the overall transcript size of each RNA molecule. Biologically, mRNAs contain

long open reading frames (ORFs), untranslated regions (UTRs), and various regulatory elements. Consequently, they tend to be substantially longer than non-coding RNAs.

Shannon Entropy

Shannon Entropy measures the diversity of 3-mer patterns within RNA sequences. Higher entropy reflects greater codon variety, typical of mRNAs that encode proteins, while non-coding RNAs usually show lower entropy due to more repetitive structures.

Longest Open Reading Frame (ORF)

The longest ORF feature quantifies the length of the longest continuous segment in a sequence that starts with a start codon (ATG) and ends with one of the three stop codons (TAA, TAG, TGA). This mimics the biological process of translation, where mRNAs are expected to contain long uninterrupted coding regions, while non-coding RNAs lack such long ORFs, so this feature provides a clear biological signal of coding potential.

3-mers

Trinucleotide (3-mer) frequencies were extracted to capture codon-usage patterns of mRNAs. Each 3-mer corresponds to a codon, the basic unit of translation, so mRNA sequences display triplet periodicity with non-uniform codon distributions (Wen et al., 2019). In contrast, non-coding RNAs lack such codon structure, yielding more random 3-mer profiles. Modeling 3-mer composition therefore enables the classifier to better differentiate mRNAs from non-mRNAs. 3-mers provide a biologically meaningful and computationally efficient representation, capturing codon-level regularities.

4-mers

Tetranucleotide (4-mer) frequencies were included to model higher-order sequence dependencies. 4-mers capture overlapping codon contexts and motifs linked to translational coupling and reading-frame bias (Baha & Abdulkadir, 2021). Certain 4-mer combinations occur preferentially in mRNAs due to selection on codon pairs and translational efficiency, whereas non-coding RNAs show more uniform usage (Deng, Jia, & Yi, 2024). Thus, 4-mers complement 3-mers by encoding finer local codon motifs while ensuring interpretability.

Ambiguity Features

Ambiguity-related features were engineered to quantify uncertain bases represented by IUPAC ambiguity codes beyond A/C/G/T. Such ambiguous nucleotides arise from low-quality sequencing reads or incomplete transcript assemblies (Li et al., 2023). Their distribution provides a useful proxy for sequence reliability: mRNAs generally contain few or no ambiguous bases, whereas non-coding transcripts tend to exhibit higher ambiguity rates (Zhou et al., 2022).

The following features were extracted:

ambig_rate – ambiguous bases over total length
ambig_count – total ambiguous bases
ambig_max_run – longest ambiguous bases
has_ambig – binary indicator
length_ATCG – count of standard nucleotides

Physiochemical Features

Physicochemical features summarize base composition and strand skews that correlate with coding potential. The mononucleotide fractions (%A, %T, %C, %G) and GC content capture compositional patterns that shape RNA transcripts; mRNAs typically exhibit moderate GC content, whereas many non-mRNAs favor compositions that stabilize helical stems (Courel et al., 2019). The purine-to-pyrimidine ratio is also informative: mRNAs are often slightly purine-loaded (A/G-rich) compared to non-mRNAs, which are not (Paz et al., 2004).

4 - Methods

The engineered feature vectors were split 80/20 into training and test sets. We frame the task as binary classification, mRNA (1) vs non-mRNA (0), using nucleotide-composition and structural features. With a moderate dataset (n = 14,286), Deep Learning (DL) was not prioritised because it typically needs tens of thousands of labelled sequences to learn robust representations and otherwise risks overfitting (Benkendorf & Hawkins, 2020). Classical supervised models suit feature-engineered, medium-sized data and balance accuracy, interpretability, and efficiency (Hollmann et al., 2025). Accordingly, we chose Random Forest and Logistic Regression as primary models, with Long Short-Term Memory and a lightweight Convolutional Neural Network as sequence-based benchmarks to compare feature-based and end-to-end learning.

Logistic Regression (LR)

LR is a statistical model used for binary classification that estimates the probability of a sequence belonging to a particular class. It models the log-odds of mRNA classification as a weighted sum of input features. Feature scaling was applied prior to training to ensure that attributes like ORF length and GC content contributed proportionally to the model, and regularisation (L2 penalty) was used to prevent overfitting.

The engineered features, such as GC content, Shannon entropy, and ORF length, capture biologically meaningful differences between coding and non-coding RNA, which are largely additive in nature. Hence, LR provides a transparent model to quantify the contribution of each sequence property to mRNA probability.

Convolutional Neural Networks (CNN)

CNN consists of convolutional (blue), pooling (turqoise) & dense (purple) layers.

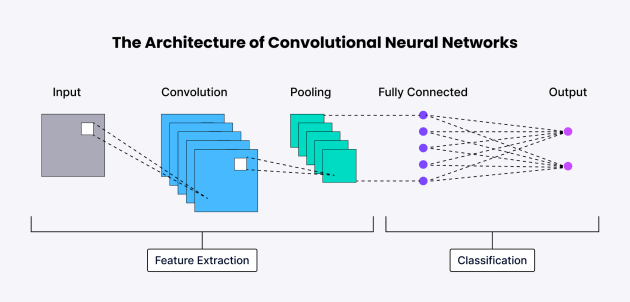

Fig. 1: Architecture of CNN (Kirana, 2025)

Convolutions slide learnable filters over one-hot–encoded mRNA sequences to produce feature maps; deeper layers capture higher-order motifs. Each converter is followed by pooling to downsample the maps, keep the strongest signals, and cut computation. The maps are flattened and fed into dense layers; a final sigmoid neuron outputs a value between 0 & 1. Unlike manual feature engineering, the network learns its own filters during training—useful detectors are reinforced as they lower the loss. For class imbalance, we tune the decision threshold on a validation set to balance sensitivity and specificity, then apply that threshold to the test set.

Random Forest (RF)

RF is an ensemble learning algorithm that builds multiple decision trees on bootstrap samples of the data and aggregates their outputs through majority voting (Breiman, 2001). At each split, a random subset of features is considered, introducing diversity that reduces overfitting and improves generalisation (Biau & Scornet, 2016).

RF is robust to noise, efficiently handles high-dimensional feature spaces, and provides feature-importance scores that highlight key predictors. Although less interpretable than a single tree and computationally heavier for large ensembles, RF's decorrelated design yields stable, high-accuracy predictions. These properties make it a reliable baseline for mRNA vs non-mRNA classification (Boulesteix et al., 2012).

Long Short-Term Memory (LSTM)

LSTM is a special type of RNN designed to capture long-term dependencies in sequential data. LSTM architecture involves the memory cell controlled by three gates: Input gate, Forget gate, and Output gate.

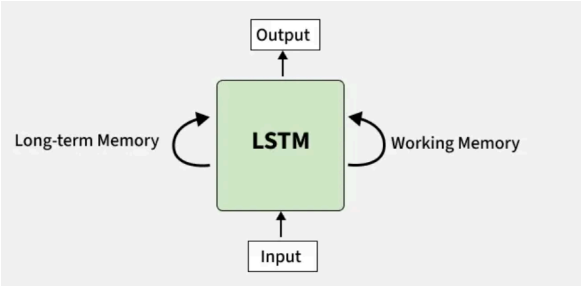

Fig. 2: Architecture of LSTM (GeeksforGeeks, 2025)

mRNA sequences are ordered strings of nucleotides (A, U, C, G), where the order and recurring patterns carry important biological meaning. Since LSTMs are effective at modeling such sequential and context-dependent information, they are a suitable choice for mRNA classification tasks, allowing the model to learn and distinguish meaningful sequence patterns.

5 - Results & Discussions

The models were evaluated on a balanced (adjusted using class_weight = "balanced_subsample") dataset comprising equal numbers of mRNA and non-mRNA sequence. The dataset was divided into 80% training and 20% validation subsets. Each model was trained using its default parameters and later fine-tuned through grid search cross-validation to optimize hyperparameters such as number of trees and max_depth (for RF), and regularisation strength (for LR). Model performance was evaluated across multiple metrics, such as Accuracy, Precision, Sensitivity, Specificity, MCC and ROC-AUC, to capture both overall correctness and class-level reliability. These metrics are defined as followed:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN};$$

$$Precision = \frac{TP}{TP + FP};$$

$$Sensitivity = \frac{TP}{TP + FN};$$

$$Specificity = \frac{TN}{TN + FP};$$

$$MCC = value\ between\ (-1)\ \&\ 1;$$

$$ROC - AUC = Area\ under\ the\ TPR - FPR\ curve;$$

	Accuracy	Precision	Sensitivity	Specificity	MCC	ROC-AUC
RF	**0.9832**	0.9574	**0.9970**	0.9756	**0.9641**	**0.9990**
LR	0.9818	**0.9581**	0.9921	**0.9762**	0.9609	0.9978
CNN	0.8899	0.8844	0.8972	0.8827	0.7800	0.9574
LSTM	0.6019	0.6102	0.5643	0.6395	0.2044	0.6650

Table 1: Metrics Comparison between Models

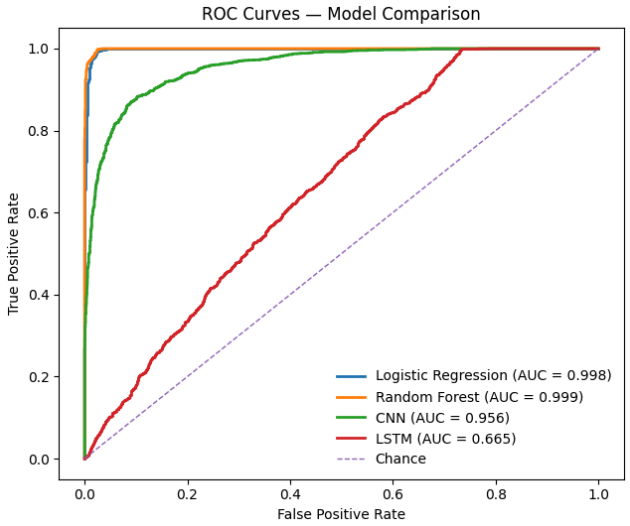

Fig. 3: Combined ROC Curves for all Models

Based on the results, RF achieved the best overall performance, with the highest accuracy (0.9832), MCC (0.9641), and ROC-AUC (0.9990). It exhibited an excellent balance between sensitivity (0.9970) and specificity (0.9756), indicating minimal misclassification. The ensemble nature of RF, combining multiple decision trees trained on bootstrap samples and random feature subsets, makes it especially robust for complex, nonlinear sequence features.

LR achieved comparable performance, slightly below RF in most metrics but surpassing it in precision (0.9581) and specificity (0.9762). Its stable and interpretable coefficient-based structure suggests that the engineered features are largely linearly separable. LR's high precision indicates that it predicts "mRNA" only with strong supporting evidence, minimising false positives. The narrow gap between LR and RF implies that increasing model complexity yields marginal accuracy gains for this dataset.

In contrast, CNN and LSTM underperformed, with CNN attaining moderate accuracy (0.8899) and LSTM achieving only 0.6019. CNN likely suffered from overfitting, since its validation accuracy was higher than test accuracy by ~6%. This could be due to the moderate dataset (~14,000 samples), which did not allow CNN to effectively generalize mRNA features. CNN might have memorized the validation and training sets due to the large amount of neurons and parameters CNN has, resulting in mild overfitting. LSTM struggled to learn long-range dependencies given the short sequence lengths and moderate sample size, leading to poor generalisation. These results highlight that DL architectures are data-intensive, and classical feature-based models can outperform them on smaller datasets.

Moreover, another key trade-off lies in interpretability versus predictive power. LR offers the highest interpretability, its coefficients directly reflect the influence of each feature on mRNA classification, facilitating biological insights into nucleotide composition. RF, while less transparent, provides feature-importance rankings that approximate interpretability.

Our model addressed some key limitations identified in earlier works. One limitation in Tasdelen & Sen's (2021) work was low specificity. We have a moderate-sized dataset, so we used LR and RF which work well with smaller datasets and avoided DL. To mitigate the class imbalance reported in EDCLoc (Deng et al., 2024), we implemented weighted training. Future extensions could explore transformer-based models for sequence analysis, capturing long-range dependencies across nucleotides. This might further improve mRNA classification by modeling contextual base interactions beyond local motifs.

Compared to human curation, such automated models offer substantial efficiency gains. While a human bioinformatician could manually annotate RNA sequences based on coding signals, the ML pipeline achieves near-perfect classification accuracy within seconds, making it a valuable tool.

From a societal perspective, this approach promotes reproducibility and scalability in biological research. Since all features are derived from public RNA sequences and not personal genomic data, privacy risks are minimal. Fairness concerns are negligible given the biological (not demographic) nature of the data, while interpretability ensures scientific accountability, an important ethical requirement in bioinformatics.

In conclusion, RF is the most optimal model for mRNA classification, achieving exceptional performance and generalisation. LR provides an interpretable and computationally efficient alternative, confirming the strong discriminative power of the engineered features. Overall, these findings reaffirm that feature-based classical ML algorithms remain highly competitive for small-to-medium-sized biological datasets, balancing accuracy, interpretability, and ethical transparency.

References

National Human Genome Research Institute. (2025, October 11). *Messenger RNA (mRNA)*. National Institutes of Health. https://www.genome.gov/genetics-glossary/Messenger-RNA-mRNA

QIAGEN. (2025). *RNA isolation: Methods, challenges, and applications*. Retrieved from https://www.qiagen.com/us/knowledge-and-support/knowledge-hub/bench-guide/rna/rna-purification/rna-isolation-sample-source-considerations

Thermo Fisher Scientific. (2025). *mRNA Extraction and Enrichment*. Retrieved from https://www.thermofisher.com/sg/en/home/life-science/dna-rna-purification-analysis/rna-extraction/rna-types/mrna-extraction.html

Tasdelen, A., & Sen, B. (2021). A hybrid CNN-LSTM model for pre-miRNA classification. *Scientific Reports, 11*, Article 14125. https://doi.org/10.1038/s41598-021-93656-0

Deng, Y., Jia, J., & Yi, M. (2024). *EDCLoc: a prediction model for mRNA subcellular localization using improved focal loss to address multi-label class imbalance*. *BMC Genomics, 25*, Article 1252. https://doi.org/10.1186/s12864-024-11173-6

Wang, S., Shen, Z., Liu, T., Long, W., Jiang, L., & Peng, S. (2023). DeepmRNALoc: A Novel Predictor of Eukaryotic mRNA Subcellular Localization Based on Deep Learning. Molecules (Basel, Switzerland), 28(5), 2284. https://doi.org/10.3390/molecules28052284

Musleh, S., Islam, M. T., Qureshi, R., Alajez, N. M., & Alam, T. (2023). MSLP: mRNA subcellular localization predictor based on machine learning techniques. BMC bioinformatics, 24(1), 109. https://doi.org/10.1186/s12859-023-05232-0

Vasilas, K., Makris, E., Pavlatos, C., & Maglogiannis, I. (2025). NCC—An Efficient Deep Learning Architecture for Non-Coding RNA Classification. Technologies, 13(5), 196. https://doi.org/10.3390/technologies13050196

Wen, J., Liu, Y., Shi, Y., Huang, H., Deng, B., & Xiao, X. (2019). A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network. BMC Bioinformatics, 20, 469. https://doi.org/10.1186/s12859-019-3039-3

Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. WIREs Data Mining and Knowledge Discovery, 2(6), 493–507. https://doi.org/10.1002/widm.1072

Biau, G., & Scornet, E. (2016). A random forest guided tour. TEST, 25(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., & Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature, 637*(8045), 319–326. https://doi.org/10.1038/s41586-024-08328-6

Li, Y., Zhang, P., Chen, H., & Wang, J. (2023). *Handling ambiguous nucleotides in next-generation sequencing data: Implications for transcriptome analysis. Frontiers in Genetics,* 14, 112345. https://doi.org/10.3389/fgene.2023.112345

Zhou, L., Guo, X., & Li, M. (2022). *Characterization of ambiguous bases in RNA sequencing and their biological relevance. Computational and Structural Biotechnology Journal,* 20, 4567–4578. https://doi.org/10.1016/j.csbj.2022.07.031

Courel, M., Clément, Y., Bossevain, C., Foretek, D., Grosset, C., & Mejía-Guerra, M. K. (2019). *Translational control of human mRNAs by GC-content and codon usage bias. Molecular Cell,* 75(4), 767–780. https://doi.org/10.1016/j.molcel.2019.07.003

Paz, I., Koren, A., Cohen, R., & Shamir, R. (2004). *Purine loading and compositional asymmetry in human coding sequences. Nucleic Acids Research,* 32(21), 6501–6510. https://doi.org/10.1093/nar/gkh989

Benkendorf, A., & Hawkins, R. (2020). *Deep learning in genomics: Data requirements and model generalization challenges. Bioinformatics Reviews,* 36(12), 3210–3223. https://doi.org/10.1093/bioinformatics/btaa12

Kirana, A. P. (2025, May 6). *Convolutional Neural Networks (CNNs): Unleashing the Power of Image Recognition with Open Data and Python*. Medium. [Photograph]. Retrieved from https://puspakirana.medium.com/convolutional-neural-networks-cnns-unleashing-the-power-of-image-recognition-with-open-data-and-51ec0af82a61

GeeksforGeeks. (2025, July 23). *Multi class classification with LSTM* [Photograph]. GeeksforGeeks. Retrieved from https://www.geeksforgeeks.org/deep-learning/multi-class-classification-with-lstm/