# A Classification Model for lncRNA and mRNA Based on K-mers and a Convolutional Neural Network

# Index


01 Introduction & Dataset

02 Preprocessing

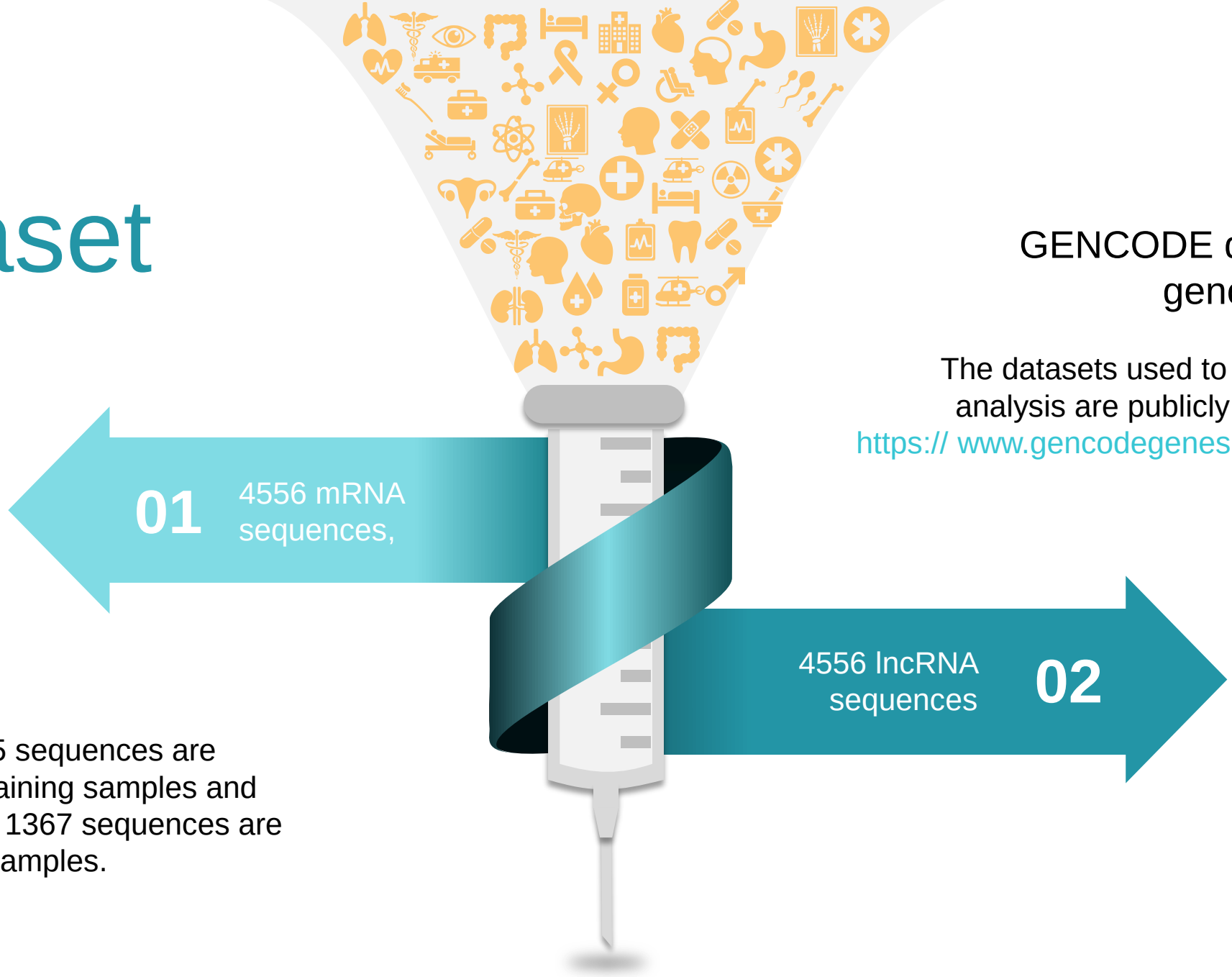
03 Activity Diagram

04 Model Convolution

05 Comparison

06 Conclusion

# Dataset

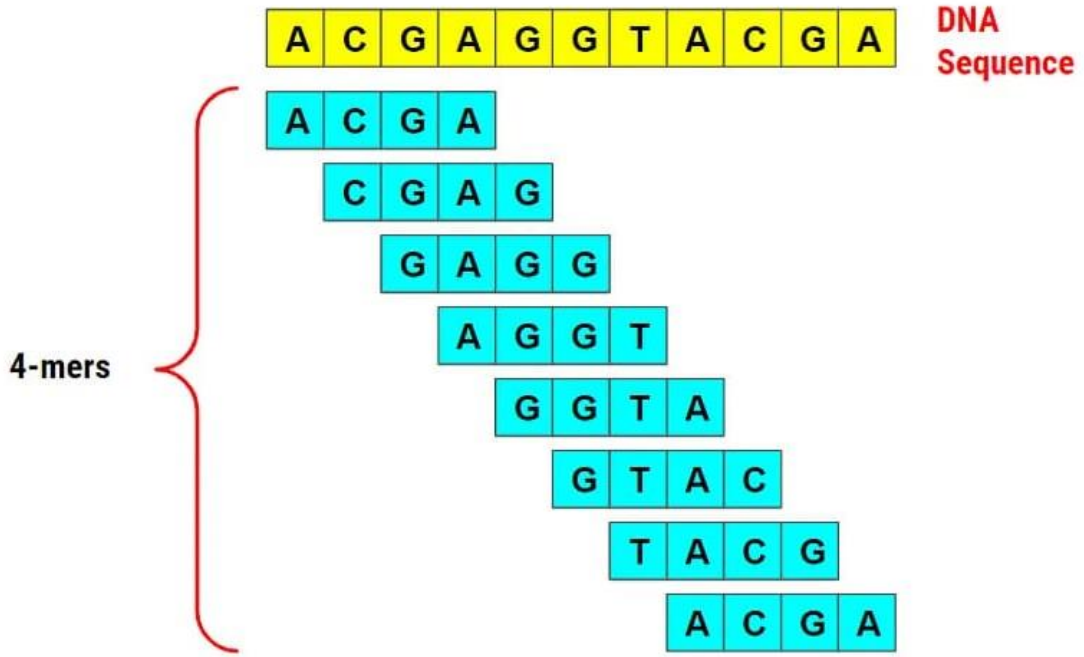



GENCODE database, gencode.v 26

The datasets used to perform the analysis are publicly available at https:// www.gencodegenes.org/mouse/

Of which 7745 sequences are selected as training samples and the remaining 1367 sequences are used as test samples.

# K-mers




K-mers are contiguous subsequences of length k that are derived from a longer sequence, typically in genomics or bioinformatics.

# Preprocessing

01

### K-mer Extraction

- Function: extract_kmers(sequence, k)
- Input: Sequence and k
- Output: List of k-mers

02

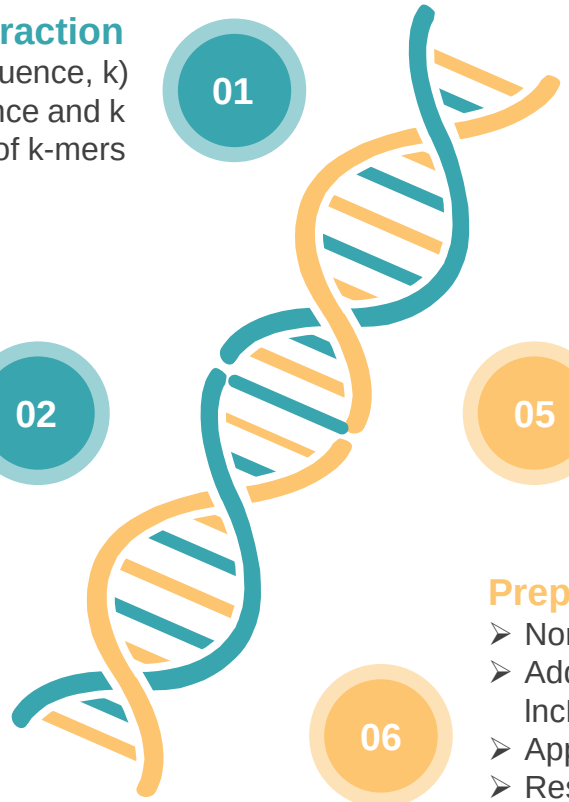
### Reading & Combining Sequence Data

- mRNA and lncRNA sequences from FASTA files using SeqIO module.
- Combining Sequences:
- Combine mRNA and lncRNA sequences into combined_sequences.

03

### K-mer Extraction and Set Creation

- Sets for unique k-mers (k=1, 2, 3, 4).
- Iterates over sequences, extracts k-mers, and updates sets.



04

### Creating K-mer Matrices

- Function: create_kmer_matrix(sequence_data, unique_kmers, k)
- Returns matrix with sequence rows, k-mer columns, and frequencies.

05

### K-mer Matrices for Different k

- Create matrices for mRNA and lncRNA (k=1, 2, 3, 4).

06

### Preprocessing for Different k:

- Normalize values in mRNA and lncRNA matrices.
- Add a column of ones for mRNA (mRNA=1) and a column of zeros for lncRNA (lncRNA=0).
- Apply preprocessing to k-mer matrices for each k (1, 2, 3, 4).
- Resulting DataFrames: df_mrna_1, df_lncRNA_1, ..., df_mrna_4, df_lncRNA_4.

# Activity Diagram

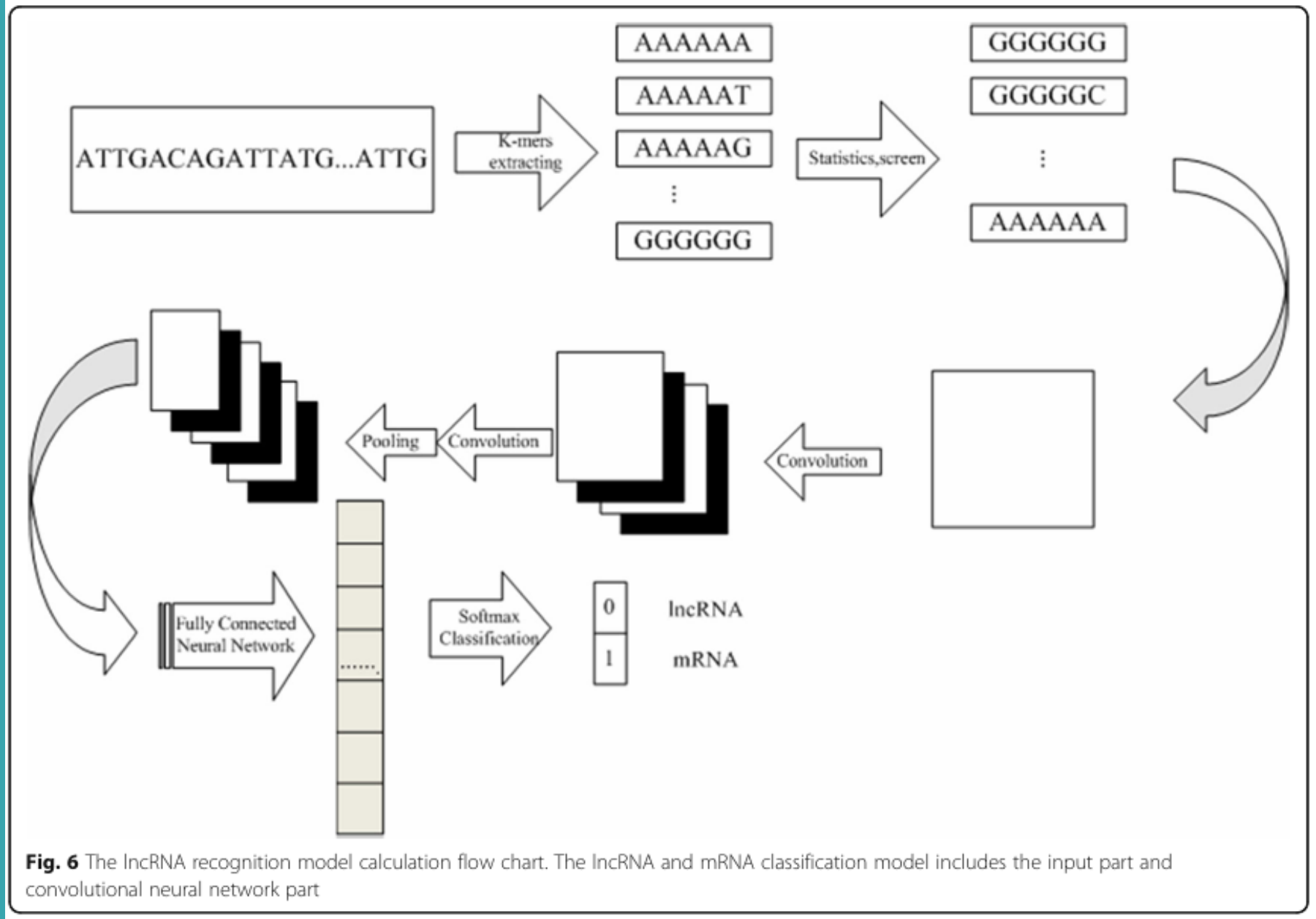
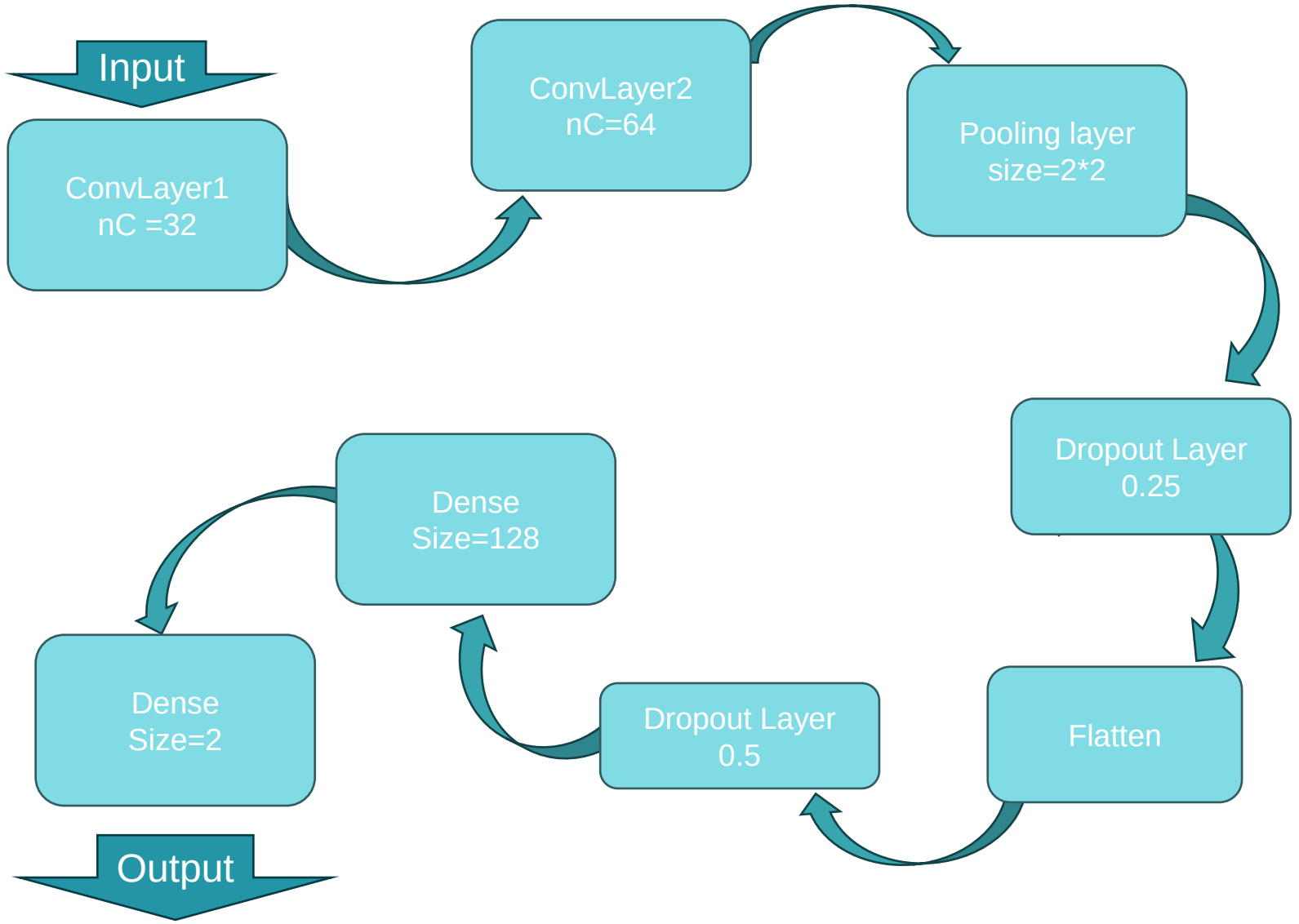
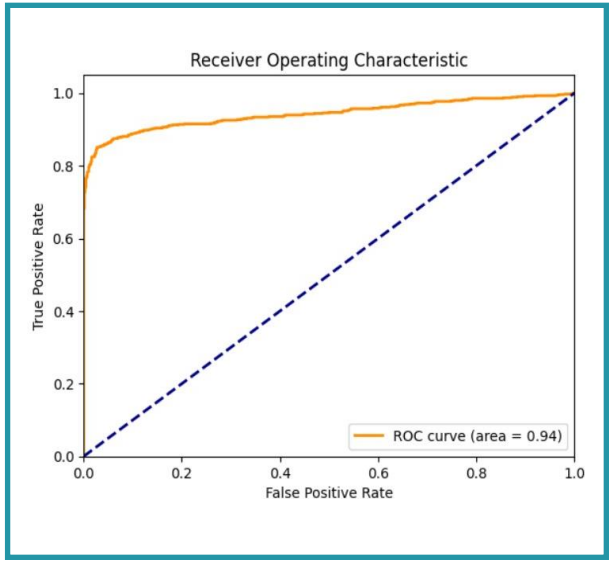



**Fig. 6** The lncRNA recognition model calculation flow chart. The lncRNA and mRNA classification model includes the input part and convolutional neural network part

# Convolution Model

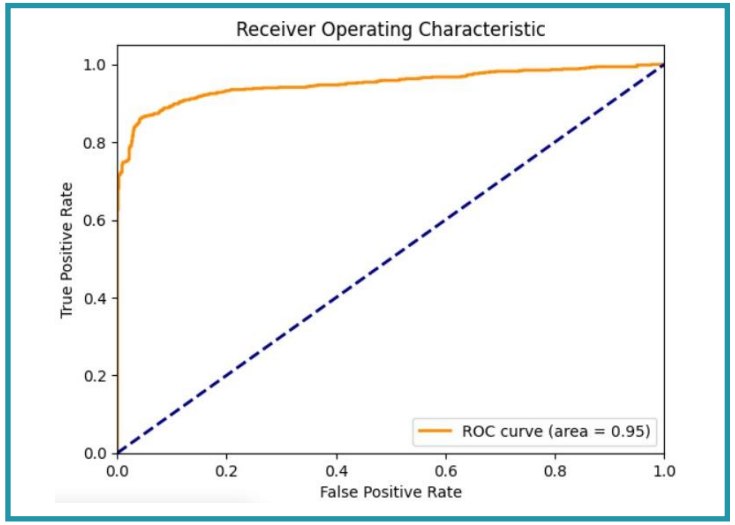


# Results and Output

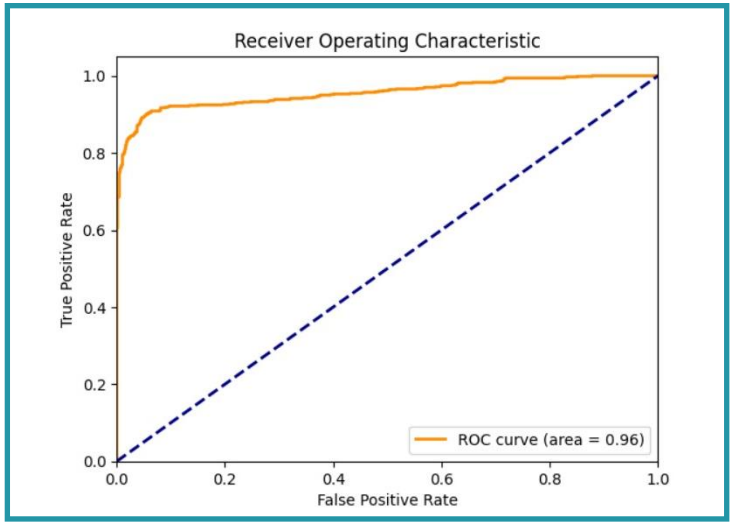



ROC curve for 3-mers (n=4^3=64)

| N( k-mers) | accuracy | Precision | Recall | F1 score |
|---|---|---|---|---|
| 16 | 0.9078 | 0.9644 | 0.8516 | 0.9045 |
| 64 | 0.9239 | 0.9771 | 0.8691 | 0.9199 |
| 256 | 0.9005 | 0.9223 | 0.8735 | 0.8972 |




ROC curve for 4-mers (n=4^4)




ROC curve for 2-mers (n=4^2)

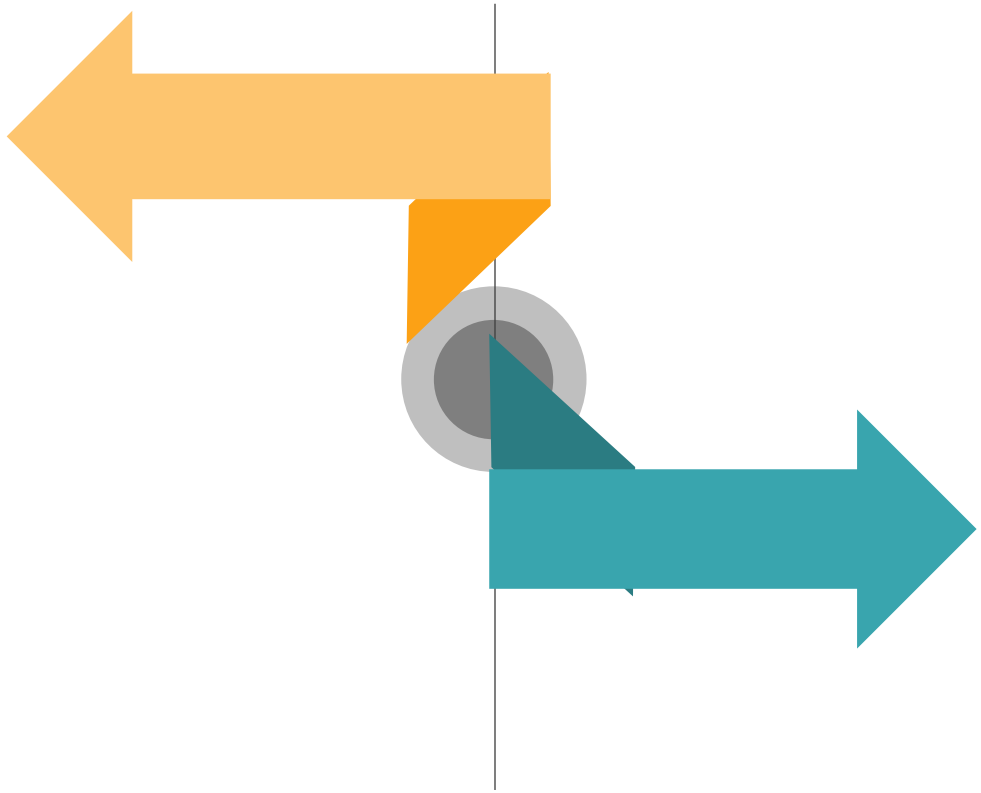
# Comparison

## CNN

Convolutional Neural Network (CNN) model can self-learn the characteristics of the sequence through continuous training **without artificial intervention and efficiently calculate large amounts of data**, no domain-expert knowledge or fine-tuning of parameters to increase accuracy are needed

**Table 5** Five model effect comparison table in human

| model | model accuracy | precision rate(P) | recall rate(R) | $F_1$score |
|---|---|---|---|---|
| CNN | 0.9872 | 0.9993 | 0.9955 | 0.9974 |
| RF | 0.8820 | 0.8949 | 0.8867 | 0.8925 |
| LR | 0.7020 | 0.7247 | 0.7183 | 0.7218 |
| DT | 0.8030 | 0.7873 | 0.7852 | 0.7869 |
| SVM | 0.7020 | 0.7245 | 0.7158 | 0.7179 |



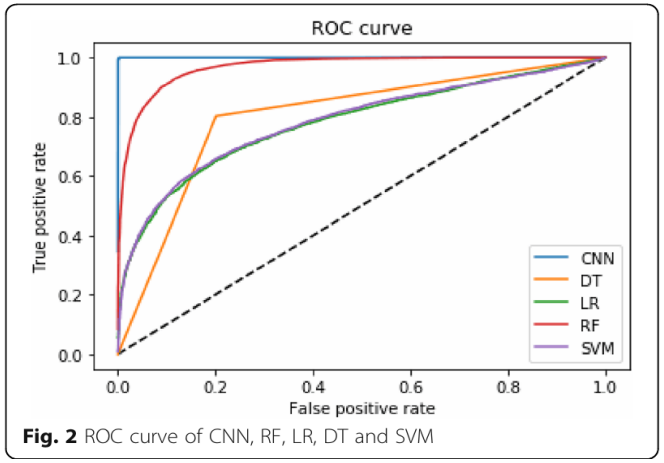



**Fig. 2** ROC curve of CNN, RF, LR, DT and SVM

## ML methods

- The study used the **maximum entropy algorithm** for k-mer screening and support vector machines for classification, revealing high computational complexity and cost.
- **Expert-driven pre-processing** and feature selection were crucial, emphasizing parameter fine-tuning for accuracy in various conventional machine learning algorithms like **SVM, logistic regression, decision trees, NN, BNs, GAs, HMMs,** etc.

# Conclusion

## Objectives Achieved

Developed an effective model for lncRNA and mRNA classification.

## Model Highlights

Utilized k-mer frequency matrix in a CNN, achieving high accuracy (0.9239) with 1-mers, 2-mers, 3-mers, and 4-mers.

## Performance Boost

Outperformed traditional models (random forest, logistic regression, decision tree, SVM) based on ROC curve analysis.

## Limitations & Future work

Acknowledged analysis simplicity, with plans for in-depth species-specific k-mer analysis. Future goal: Develop a user-friendly web server.

## Conclusion

This study not only met its classification objectives but also identified avenues for improvement and broader applicability in bioinformatics.

# Thank You