mathematics

MDPI

Article

LightGBM-LncLoc: A LightGBM-Based Computational Predictor for Recognizing Long Non-Coding RNA Subcellular Localization

Jianyi Lyu, Peijie Zheng, Yue Qi and Guohua Huang *

School of Information Engineering, Shaoyang University, Shaoyang 422000, China
* Correspondence: 3280@hnsyu.edu.cn or guohuahhn@163.com

Abstract: Long non-coding RNAs (lncRNA) are a class of RNA transcripts with more than 200 nucleotide residues. LncRNAs play versatile roles in cellular processes and are thus becoming a hot topic in the field of biomedicine. The function of lncRNAs was discovered to be closely associated with subcellular localization. Although many methods have been developed to identify the subcellular localization of lncRNAs, there still is much room for improvement. Herein, we present a lightGBM-based computational predictor for recognizing lncRNA subcellular localization, which is called LightGBM-LncLoc. LightGBM-LncLoc uses reverse complement k-mer and position-specific trinucleotide propensity based on the single strand for multi-class sequences to encode LncRNAs and employs LightGBM as the learning algorithm. LightGBM-LncLoc reaches state-of-the-art performance by five-fold cross-validation and independent test over the datasets of five categories of lncRNA subcellular localization. We also implemented LightGBM-LncLoc as a user-friendly web server.

Keywords: lncRNA; subcellular localization; lightGBM; reverse complement k-mer; machine learning

MSC: 92B15; 68T99

Citation: Lyu, J.; Zheng, P.; Qi, Y.; Huang, G. LightGBM-LncLoc: A LightGBM-Based Computational Predictor for Recognizing Long Non-Coding RNA Subcellular Localization. *Mathematics* **2023**, *11*, 602. https://doi.org/10.3390/math11030602

Academic Editor: Tianhai Tian

Received: 15 September 2022
Revised: 22 December 2022
Accepted: 5 January 2023
Published: 25 January 2023

1. Introduction

Long non-coding RNAs (lncRNAs) generally refer to a type of RNA transcripts with more than 200 nucleotides which are transcribed from DNA but never code for proteins [1,2]. A large body of evidence indicates that lncRNAs act as key regulators by binding RNA, DNA, or proteins in numerous cellular processes, including the cell cycle [3], differentiation [4,5], and metabolism [6,7]. For example, some lncRNAs control gene expression; some affect replication or the response to DNA damage and repair; some are involved in splicing, turnover, translation, and signal pathways [8]; and some lncRNAs target miRNAs and mRNAs, so their localization is very important [9]. Xing et al. discovered the involvement of lung cancer-related transcript 1 (LUCAT1) in the regulation of multiple tumors, including lung cancer, breast cancer, ovarian cancer, thyroid cancer, and renal cell carcinoma; thus, LUCAT1 was viewed as a potential prognostic biological marker and therapeutic target for cancer [10]. The roles of lncRNAs in cellular processes are closely associated with their subcellular localization [11], which determines which partners they interact with, as well as what post- or co-transcriptional regulatory modifications occur, and influence the external stimuli directly impacting lncRNA function [12]. For example, nuclear lncRNAs are, overall, more abundant and less stable than cytoplasm lncRNAs, so the nuclear lncRNAs function differently from cytoplasm lncRNAs. The former modulate transcriptional programs through chromatin interactions and remodeling [13–15], while the latter mediate signal transduction pathways, translational programs, and posttranscriptional control of gene expression [12]. Therefore, knowledge about lncRNA subcellular localization is helpful to infer or understand its functions.

Benefiting from advances in artificial intelligence, no less than ten computational methods have been developed in the past two decades for predicting the subcellular localization of protein [16–20]. For instance, Hua et al. [21] proposed a support vector machine (SVM)-based tool for recognizing the subcellular localization of proteins both in prokaryotic organisms and in eukaryotic organisms in the early year of 2001, and Almagro Armenteros et al. [22] developed a recurrent neural network and attention-based deep learning method for predicting protein subcellular localization. Shen et al. [23] comprehensively compared and evaluated these developed web-based tools for predicting human protein subcellular localization across various benchmark datasets. However, the prediction of lncRNA subcellular localization lags seriously behind the prediction of protein subcellular localization. Until the last 5 years, only a few computational methods had been proposed to address lncRNA subcellular localization prediction. Cao et al. [24] developed an ensemble-learning-based method (lncLocator) to predict five types of lncRNA cellular localization. Cao et al. [24] extracted k-mer features and high-level features from primary sequences and used them to train two SVM classifiers and two random forest classifiers, respectively. The four classifiers were stacked to determine lncRNA subcellular localization. Su et al. [25] used the general pseudo-k-tuple nucleotide composition (PseKNC) to represent lncRNA sequences and built an SVM-based method (called iLoc-lncRNA) for lncRNA subcellular localization prediction. Because the numbers of lncRNAs located in different organelles are imbalanced, models built on the imbalanced datasets would be in favor of categories with dominating lncRNAs in terms of number. Feng et al. [26] extracted multi-source representations from multiple perspectives and employed heterogeneous fusion and feature selection to overcome the impact of data imbalance. Gudenas et al. [27] used multiple informative features including k-mers to represent lncRNAs and developed a deep neural-network-based method for the prediction of lncRNA subcellular localization. Zeng et al. [28] suggested that k-mers are the frequencies of different continuous nucleotide combinations that are sufficient to extract simple sequence motifs. In fact, k-mers lose information about their order and position and thus fail to extract semantic relationships between nucleotides or nucleotide combinations in the context. Zeng et al. [28] used sequence semantics captured by word2vec [29,30] in addition to k-mers to improve the representation of lncRNAs. For lncRNA sequences of variable length, it is very inconvenient to extract features from the whole sequence. Inspired by the idea of spatial pyramid pooling [31], Zeng et al. [28] divided lncRNA sequences into m consecutive subsequences, each represented by semantics, as well as k-mers. The average pooling of m representations was the representation of a lncRNA. However, average pooling is not an optimal strategy: it averaged representations over m subsequences, which lost remarkable representations. Here, we used the reverse complement k-mer (RCKmer), as well as the position-specific trinucleotide propensity based on a single strand for multi-class sequences (PSTNPSSMC), to represent lncRNA, and we propose a Light Gradient Boosting Machine (LightGBM)-based method for classifying five types of lncRNA subcellular localizations.

2. Methods

2.1. Dataset

The training dataset was downloaded directly from DeepLncLoc [28] (http://bioinformatics.csu.edu.cn/DeepLncLoc/ accessed on 16 December 2021), which is a newly developed method for predicting lncRNA subcellular localization. The original data were retrieved from the RNALocate database [32], which contains 42,190 manually curated and experimentally validated RNA-associated subcellular localization entries. Zeng et al. [28] preprocessed these data by the following steps: (a) lncRNAs with multiple entities were merged into a unique entity according to the identical gene name, (b) lncRNAs with multiple localizations were removed, and (c) categories with fewer than 10 lncRNAs were excluded. Five subcellular localization entities (nucleus, cytosol, cytoplasm, ribosome, and exosome) were finally preserved, containing 857 lncRNAs. The numbers of lncRNAs with subcellular localization in the nucleus, cytosol, cytoplasm, ribosome, and exosome were 325, 88, 328, 88, and 28,

respectively. The maximum length of the lncRNA sequences was 55,120, the minimum was 166, and the average was 8216.

The independent test dataset was also downloaded from DeepLncLoc [28]. LncRNAs of three types of subcellular localization (nucleus, cytoplasm, and ribosome) were obtained from the lncSLdb database [33], and lncRNA of the two other types of subcellular localization (cytosol and exosome) were retrieved by Zeng et al. [28] from the medical literature database PubMed (https://pubmed.ncbi.nlm.nih.gov/ accessed on 25 December 2021) with the combined keywords lncRNA and cytosol or lncRNA and exosome. All the lncRNA sequences were retrieved from the NCBI database, and sequence redundancies were decreased to 0.9 using CD-Hit [34,35], a commonly used sequence clustering program. The independent test data finally consisted of 67 lncRNAs, of which 20 belonged to the cytoplasm, 20 to the nucleus, 10 to the ribosome, 10 to the cytosol, and 7 to the exosome.

2.2. Methodology

As shown in Figure 1, the proposed method was composed of sequence division, representation of sequences, training the classifier, and predicting the subcellular localization of lncRNAs. The lncRNAs differ greatly in sequence length. It is impossible for machine learning algorithms to deal with these lncRNAs of variable length. Therefore, the first step is to process lncRNA sequences to a fixed length. We cut the first 166 nucleotide residues from each lncRNA sequence. Then, we computed the PSTNPSSMC and RCKmer [36,37] of lncRNA subsequences and employed LightGBM [38,39] to learn a classifier. The trained LightGBM finally predicted the subcellular localization of the unknown lncRNAs that had been preprocessed and then represented by the PSNTPSSMC and RCKmer.

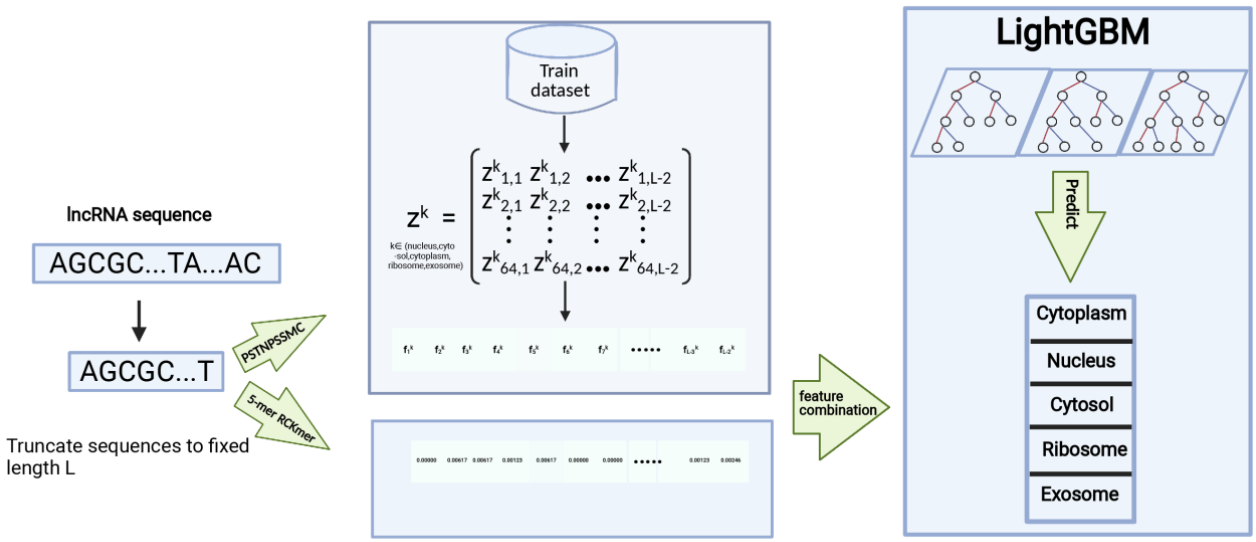

Figure 1. The overview of the proposed LightGBM-LncLoc.

2.3. Feature Selection

2.3.1. PSTNPSSMC

The position-specific trinucleotide propensity based on the single strand (PSTNPSS) was initially designed to extract features from binary-class sequences [40]. Here, we extended it to deal with multi-class sequences. Assume that the lncRNA sequence S is $s_1 s_2 \cdots s_L$, where L is the uniform length of S. The lncRNA sequence S is then divided into 3-mers in one stride, namely, $s_1 s_2 s_3$, $s_2 s_3 s_4$, $\cdots$, $s_i s_{i+1} s_{i+2}$, $\cdots$, $s_{L-2} s_{L-1} s_L$.

Moreover, $s_i s_{i+1} s_{i+2}$ is one of 64 types of 3-mer nucleotides, where i denotes the position in the sequence. The position-specific trinucleotide propensity matrix is denoted by

$$Z^k = \begin{bmatrix} z^k_{1,1} & z^k_{1,2} & \cdots & z^k_{1,\,L-2} \\ z^k_{2,1} & z^k_{2,2} & \cdots & z^k_{2,L-2} \\ \vdots & \vdots & \vdots & \vdots \\ z^k_{64,1} & z^k_{64,2} & \cdots & z^k_{64,L-2} \end{bmatrix} \tag{1}$$

where $z^k_{i,j}$ denotes the occurrence probability of the i-th type of 3-mer at the j-th position in the sequence, and k $\in$ {nucleus, cytosol, cytoplasm, ribosome, exosome}. $z^k_{i,j}$ can be estimated from the corresponding occurrence frequency in all the samples of the training dataset. For each $s_i s_{i+1} s_{i+2}$, we consult the position-specific trinucleotide propensity matrix Z^k to obtain its frequency f^k_i. f^k_i is sorted from large to small at the given position i. We use a label array C of the same size as f^k_i to record the type of the subcellular localization. We count the occurrence numbers of five types of subcellular localization in the array. The position-specific trinucleotide propensity matrix with the maximum occurrence number among the five types is used to characterize the lncRNA sequence. For example, if the cytosol occurs the most times in the label array, the lncRNA sequence S is represented by

$$S = \left(f^{cytosol}_1, f^{cytosol}_2, \ldots, f^{cytosol}_{L-2}\right) \tag{2}$$

where $f^{cytosol}_i$ belongs to the set of elements in the i-th column of $Z^{cytosol}$. Let us take a simple sequence, AACGCCT, as an example. The types of 3-mers AAC, ACG, CGC, GCC, and CCT were assumed to correspond to 2, 7, 26, 38, and 24, respectively. Then, we compare all $z^k_{2,1}$ k $\in$ {nucleus, cytosol, cytoplasm, ribosome, exosome} and choose the max number and record its localization type k in label array C. Next, we compare $z^k_{7,2}$, $z^k_{26,3}$, $z^k_{38,4}$, and $z^k_{24,5}$ in order. Assuming that the label array of sequence AACGCCT is C[cytosol, cytoplasm, ribosome, exosome, cytosol], it is not hard to see that the maximum localization type is cytosol. Then, we use all values from $Z^{cytosol}$, that is, $(z^{cytosol}_{2,1}, z^{cytosol}_{7,2}, z^{cytosol}_{,26,3}, z^{cytosol}_{38,4}, z^{cytosol}_{24,5})$, as the feature encoding.

2.3.2. RCKmer

The RCKmer [36,37] is a variant of k-mer representation. The reverse complement of a DNA sequence is organized by exchanging T and A and exchanging G and C in the original sequences, i.e., A and T are paired, and G and C are paired [41]. The RCKmer views two complemented k-mer nucleotides as identical k-mers. For example, in the RCKmer, 'TT' is identical to 'AA' and 'GG' is identical to 'CC'. For k = 2, there are 16 types of traditional k-mer, but there are only 10 types of RCKmer, i.e., 'AA', 'AC', 'AG', 'AT', 'CA', 'CC', 'CG', 'GA', 'GC', and 'TA'. The number of RCKmer types is calculated by

$$\begin{cases} 2^{2k-1} & k = 1, 3, 5, \cdots \\ 2^{2k-1} + 2^{k-1} & k = 2, 4, 6, \cdots \end{cases} \tag{3}$$

2.4. LightGBM

LightGBM [38] is a highly efficient implementation of the gradient boosting decision tree (GBDT) [42], which is a popular machine learning algorithm. The GBDT is an additional model. At each iteration, it learns a new decision tree that is intended to fit residues between the target and sum over outputs of the previous decision. In the process of implementing the GBDT, there are a few schemes to solve the objective, including the exact greedy strategy and approximate algorithms. Even for the same scheme, there are still different physical implementations. The well-known XGBoost [43], pGBRT [44], scikit-learn [45], and gbm in R [46] are representatives of these implementations. The lightGBM uses Gradient-based One-Side Sampling [38] (GOSS) to reduce the number of training samples and Exclusive

Feature Bundling [47] (EFB) to bundle features. GOSS assumes that the samples with larger gradients contribute more to the information gain than do the ones with smaller gradients. GOSS first sorts all the samples by the absolute value of their gradients in descending order and then chooses a certain proportion of top samples. GOSS also samples a certain proportion of samples from the remaining data. EFB models the bundling of features as the graph coloring where it views the features as vertices and the conflict between features as weighted edges. Both GOSS and EFB greatly reduce the computing complexity, but not at the expense of the predictive accuracy.

2.5. Validation and Metrics

We utilized two ways to check the performance of the proposed method. One is fivefold cross-validation, and the other is the independent test. Five-fold cross-validation divides the training dataset into five parts, of which four parts are used to train the model and one is used to test the method. The process is repeated five times. The independent test uses data completely different from the training dataset to test the trained method. The accuracy (ACC), Macro F-measure, and area under the receiver operator characteristic curve (AUC) were adopted to evaluate the predictive results. The ACC is calculated by

$$\mathrm{ACC} = \frac{\sum_{i=1}^{N} \mathrm{I}(\mathrm{pred}_i == \mathrm{label}_i)}{\mathrm{N}}, \tag{4}$$

where N is the total number of all the samples and I is an indicator function.

The Macro F-measure is computed by

$$\mathrm{Macro\ F-measure} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{2{*}\mathrm{precision}^{(i)}{*}\mathrm{recall}^{(i)}}{\mathrm{precision}^{(i)} + \mathrm{recall}^{(i)}} \tag{5}$$

where

$$\mathrm{Precision}^{(i)} = \frac{\mathrm{TP}^{(i)}}{\mathrm{TP}^{(i)} + \mathrm{FP}^{(i)}} \tag{6}$$

$$\mathrm{Recall}^{(i)} = \frac{\mathrm{TP}^{(i)}}{\mathrm{TP}^{(i)} + \mathrm{FN}^{(i)}} \tag{7}$$

$$\mathrm{Macro\ Precision} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{\mathrm{TP}^{(i)}}{\mathrm{TP}^{(i)} + \mathrm{FP}^{(i)}} \tag{8}$$

$$\mathrm{Macro\ Recall} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{\mathrm{TP}^{(i)}}{\mathrm{TP}^{(i)} + \mathrm{FN}^{(i)}} \tag{9}$$

Although this is a multi-class question, it is feasible to calculate the Precision and Recall as for a binary problem. When calculating the Precision and Recall of the i-th category, all the samples of this category were viewed as positive, and others were viewed as negative. Therefore, $\mathrm{TP}^{(i)}$, $\mathrm{FN}^{(i)}$, and $\mathrm{FP}^{(i)}$ denote the numbers of true positive samples, false negative samples, and false positive samples, respectively, for the i-th category.

3. Results and Discussion

3.1. Feature Combination Optimization

In the field of machine learning, representations of samples play equivalent roles in predictive accuracy with learning algorithms. Therefore, developing informative representations is also what is firstly emphasized in recognizing long non-coding RNA subcellular localization. More than ten representations have been developed to represent RNA sequences over the past decades, such as the frequently used composition of k-spaced nucleic acid pairs (CKSNAP) [48], nucleotide chemical properties (NCP) [49], accumulated nucleotide frequency (ANF) [36], enhanced nucleic acid composition (ENAC) [36], and nucleic acid composition (NAC) [36]. We conducted five-fold cross-validation over the training

dataset to compare these representations. The performance evaluation is presented in Table 1. The PSTNPSSMC performed best, reaching an average ACC of 0.673, an average AUC of 0.895, and a Macro-F-measure of 0.290. The RCKmer was second only to the PSTNPSSMC, and the ANF performed worst.

Table 1. Performance using LightGBM on a single representation under 5-fold cross-validation.

Feature Type	ACC	AUC	Macro F-Measure
PSTNPSSMC	0.673	0.895	0.776
RCKmer	0.512	0.616	0.290
CKSNAP	0.467	0.592	0.267
NAC	0.466	0.590	0.240
ENAC	0.463	0.612	0.261
NCP	0.454	0.591	0.246
ANF	0.429	0.531	0.233

We used the forward-searching strategy to look for an optimal combination of representations. The forward-searching strategy first sorted all the representations by the predictive ACC in descending order and then chose the best single representation as the initial expanding set. As shown in Table 1, the best single representation was PSTNPSSMC, followed successively by RCKmer, CKSNAP, NAC, ENAC, NCP, and ANF. Next, the expanding set was combined with other single representations, i.e., PSTNPSSMC was combined respectively with RCKmer, CKSNAP, NAC, ENAC, NCP, and ANF. The combinations better than the best single representation were reserved to update the expanding set. We continued to combine the expanding set with other single representations, and the better combinations were preserved to update the expanding set. The process was repeated until the combinations did not bring better prediction accuracy than using only PSTNPSSMC. Table 2 lists all combinations of features for which the accuracy was greater than that of PSTNPSSMC alone. The combination of PSTNPSSMC and RCKmer reached the best ACC, so we chose PSTNPSSMC and RCKmer as the lncRNA representation.

Table 2. Performance using LightGBM on combinations of different representations under 5-fold cross-validation.

Feature Type	ACC	AUC	Macro F-Measure
PSTNPSSMC	0.673	0.895	0.776
PSTNPSSMC + RCKmer	0.696	0.904	0.772
PSTNPSSMC + CKSNAP	0.687	0.899	0.758
PSTNPSSMC + RCKmer + CKSNAP	0.695	0.906	0.775
PSTNPSSMC + RCKmer + NAC	0.686	0.901	0.769
PSTNPSSMC + CKSNAP + NAC	0.678	0.905	0.780

3.2. Selection of Learning Algorithms

The predictive performance also depends on the learning algorithms. We tested six popular machine learning algorithms for classifying the subcellular localization of lncRNA, i.e., LightGBM [38], XGBoost [50], support vector machine [51], random forest [52], logistic regression [53], and multilayer perceptron [54]. We used these six machine learning algorithms with the seven common feature encodings described above under five-fold cross-validation. The performance results are provided in Supplementary Tables S1–S5. Except for the logistic regression algorithm, the algorithms had the best results on the feature PSTNPSSMC. The predictive accuracy of the six algorithms under five-fold cross-

validation with feature PSTNPSSMC is presented in Table 3. LightGBM performed best, reaching an accuracy of 0.673, which was more 0.028 than the second-best. Therefore, we chose LightGBM as the learning algorithm for classification of the subcellular localization of lncRNA.

Table 3. Performance of different machine learning algorithms with feature PSTNPSSMC under 5-fold cross-validation.

Methods	ACC	AUC	Macro F-Measure
LightGBM	0.673	0.895	0.776
XGBoost	0.645	0.895	0.735
SVM	0.602	0.882	0.700
Random Forest	0.632	0.890	0.752
Logical regression	0.395	0.839	0.242
Multilayer perceptron	0.557	0.854	0.664

3.3. Parameter Optimization

In both LightGBM and RCKmer, there exist some parameters which would influence the predictive accuracy. In order to improve the predictive performance, we further optimized these parameters. In the feature PSTNPSSMC combined with RCKmer, the parameter k of the RCKmer denotes the number of continuous nucleotide resides. We investigated the effect of different k values on performance. Table 4 presents the predictive accuracies under five-fold cross-validation for k = 2, 3, 4, 5, and 6. With the increment in k, the accuracy increased slowly. When k was 5, the accuracy reached the best value of 0.700, and then it decreased dramatically. Therefore, we set k to 5 in the subsequent experiments.

Table 4. Performance of feature PSTNPSSMC combined with various values of parameter k in the RCKmer.

k	ACC	AUC	Macro F-Measure
2	0.692	0.906	0.785
3	0.696	0.904	0.772
4	0.698	0.903	0.793
5	0.700	0.905	0.779
6	0.672	0.897	0.759

The lightGBM has many parameters to set, such as the minimal data in one leaf node (min_data_leaf), feature fraction (feature_fraction), maximum depth in one tree (max_depth), and maximum number of leaves in one tree (num_leaves). These parameters affect its performance. As shown in Figure 2, the predictive performance varied greatly with the parameter values. The accuracy curve fluctuated as the minimal data in one leaf node ranged from 10 to 30. The accuracy reached the highest performance at 17 (Figure 2a). The accuracy first decreased dramatically as the feature fraction increased and then ascended dramatically to the best at feature fraction 1 (Figure 2b). The accuracy fluctuated with the maximum depth in one tree, reaching the best when it was 5 (Figure 2c). The accuracy fluctuated dramatically with the maximum number of leaves in one tree and then tended to be stable when the maximum number of leaves was 24 (Figure 2d). Therefore, we set the minimal data in one leaf node, the feature fraction, the maximum depth in one tree, and the maximum number of leaves in one tree to 17, 1, 5, and 24, respectively.

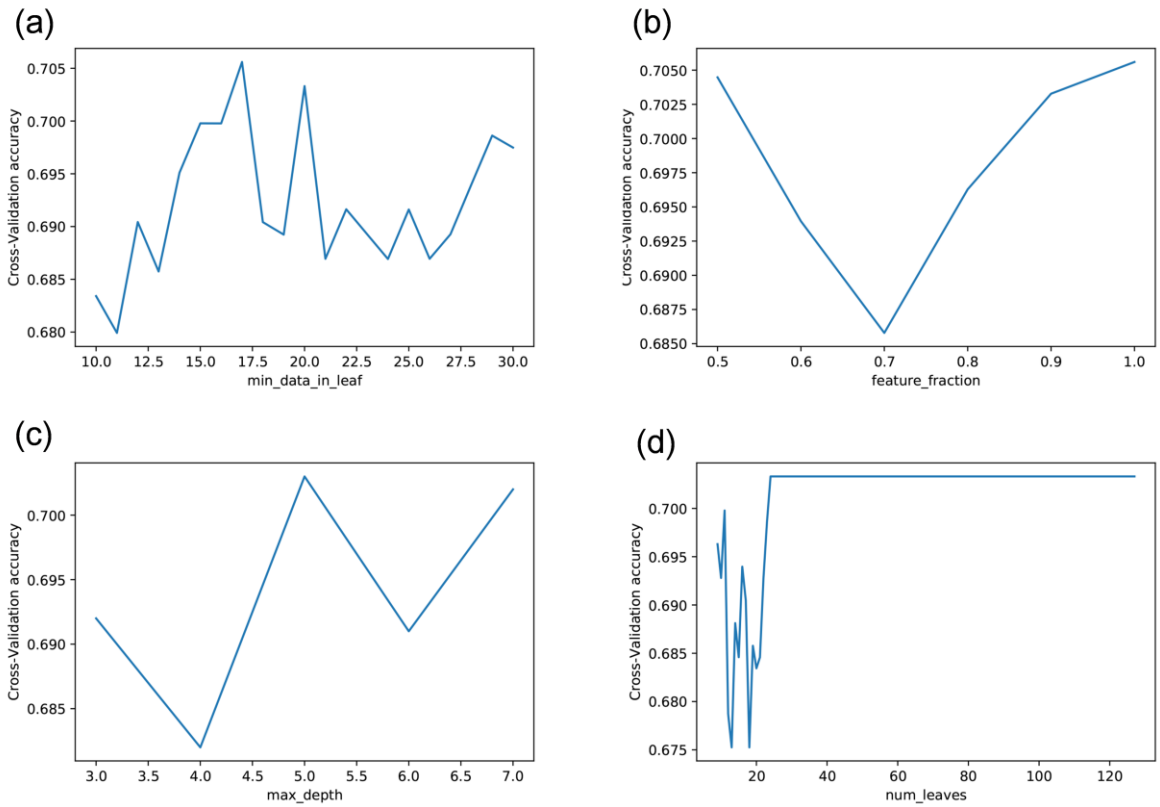

Figure 2. Effect of adjusting different parameters of LightGBM on accuracy: (**a**) min_data_in_leaf, (**b**) feature_fraction, (**c**) max_depth, (**d**) num_leaves.

LncRNAs are of variable length. It is a challenging task for machine learning algorithms to deal with variable-length sequences. We simply cut the first 166 nucleotide residues of sequences to represent lncRNAs; 166 was the minimum length of all the training lncRNA sequences. We compared this choice with the last 166 nucleotide residues and continuous 166 nucleotide residues at any position. Table 5 lists their performance after five-fold cross-validation. The first 166 nucleotide residues more informatively reflected the type of subcellular localization than did the last 166 nucleotide residues, followed by 166 continuous nucleotide residues at any position.

Table 5. Comparison of lncRNA subsequences at various positions.

Location	ACC	AUC	Macro F-Measure
The first 166 of lncRNAs	0.703	0.904	0.792
The last 166 of lncRNAs	0.607	0.887	0.704
Random 166 of lncRNAs	0.583	0.862	0.671

3.4. Comparison with State-of-the-Art Methods

Over the past ten years, no less than ten computational methods have been developed for distinguishing or recognizing the subcellular localization of lncRNAs. Some methods have not been developed into tools. Some provided stand-alone software or web servers, but they did not work or are not available at present. Based on these considerations, we compared our proposed method with only two state-of-the-art methods: iLoc-lncRNA [25] and DeepLncLoc [28]. iLoc-lncRNA can predict four types of subcellular localizations (nucleus, cytoplasm, ribosome, and exosome). DeepLncLoc can predict five types of subcellular localizations (nucleus, cytoplasm, ribosome, cytosol, and exosome). We used their web servers to compare the predictions (iLoc-lncRNA is available at http://www.csbio.sjtu.edu.cn/bioinf/lncLocator/ (accessed on 13 April 2022) and

DeepLncloc is available at http://bioinformatics.csu.edu.cn/DeepLncLoc/ accessed on 13 April 2022). To the best of our knowledge, DeepLncLoc is the latest method to classify the multi-type subcellular localization of lncRNAs. Tables 6 and 7 present the performance of five-fold cross-validation and the independent test, respectively. Obviously, both on the cross-validation and on the independent test, LightGBM-LncLoc completely outperformed DeepLncLoc. For example, LightGBM-LncLoc increased the ACC by 0.158 over DeepLncLoc on cross-validation, and it increased the ACC by 0.03 over DeepLncLoc and by 0.06 over iLoc-lncRNA on the independent test. We conducted five-fold cross-validation 10 times. A box plot is shown in Figure 3. The average ACC is still 0.698, far exceeding that of DeepLncLoc. Table S8 shows the details of these ten experiments. Figure 4 shows the ROC curves of five categories of lncRNA subcellular localization over the five-fold cross-validation. It was observed that the performance of each class was different. The performance was better for the cytosol, exosome, and ribosome than for the other two categories. Table 8 presents the performance of each category of lncRNA subcellular localization over the independent test. iLoc-LncRNA cannot predict the cytosol subcellular localization type. In order to compare the performance objectively and fairly, all cytosol sequences in the test set were removed when testing iLoc-LncRNA. There was a difference between the predictive accuracies of the lncRNA categories. For example, LightGBM-LncLoc reached the best Recall (0.900) on the nucleus and reached the worst Recall (0.400) on the cytoplasm and ribosome. The best Precision was 1 on the cytosol and exosome, while the worst was 0.429 on the nucleus. As shown in Table 8, except for exosome, LightGBM-LncLoc outperformed DeepLncLoc in terms of F-measure, indicating the superiority of LightGBM-LncLoc. Figures S1 and S2 show the confusion matrices in the independent test with DeepLncLoc and iLoc-lncRNA.

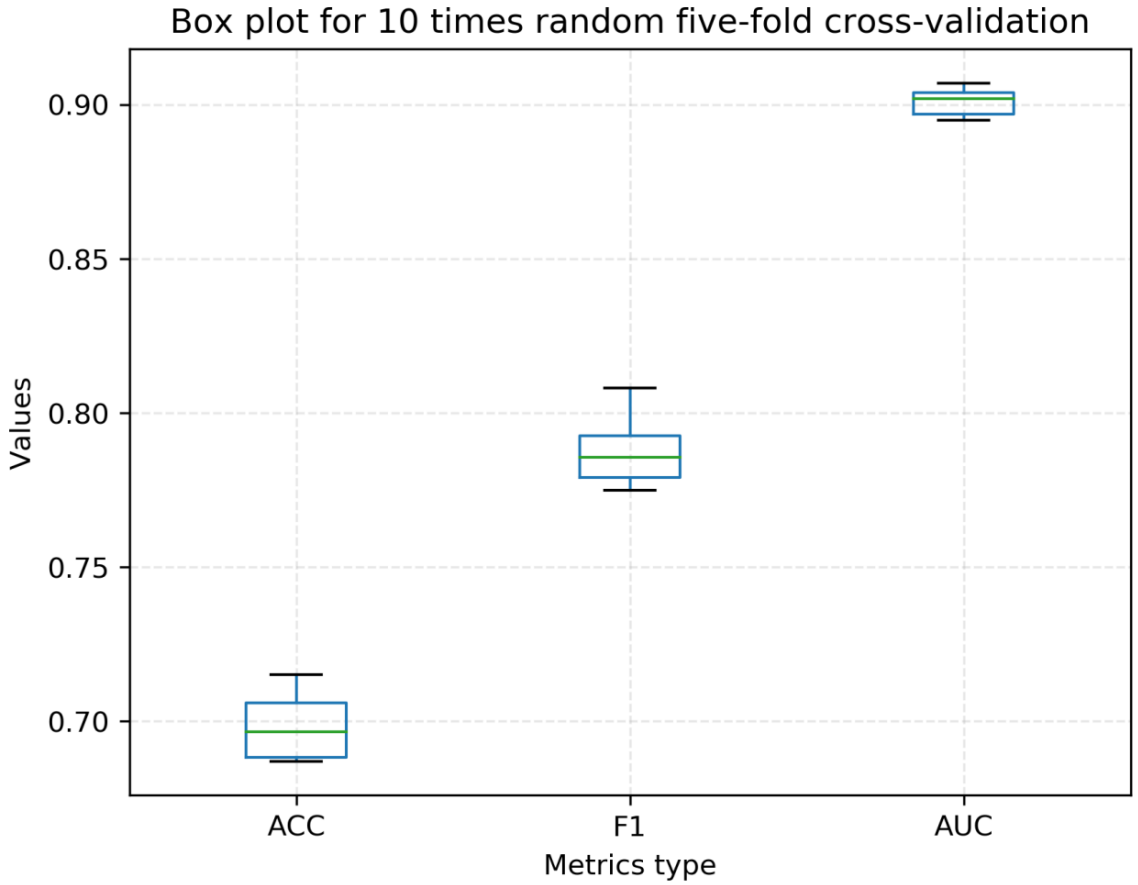

Figure 3. Box plot of LightGBM-Lncloc with 10 random 5-fold cross-validations.

Table 6. Performance on 5-fold cross-validation.

Predictor	ACC	Macro F-Measure	AUC
DeepLncLoc	0.548	0.420	0.820
LightGBM-LncLoc	0.706	0.786	0.904

Table 7. Performance on the independent test.

Predictor	Macro Precision	Macro Recall	Macro F-Measure	ACC
iLoc-lncRNA	0.488	0.445	0.458	0.507
DeepLncLoc	0.702	0.524	0.563	0.537
LightGBM-LncLoc	0.779	0.525	0.576	0.567

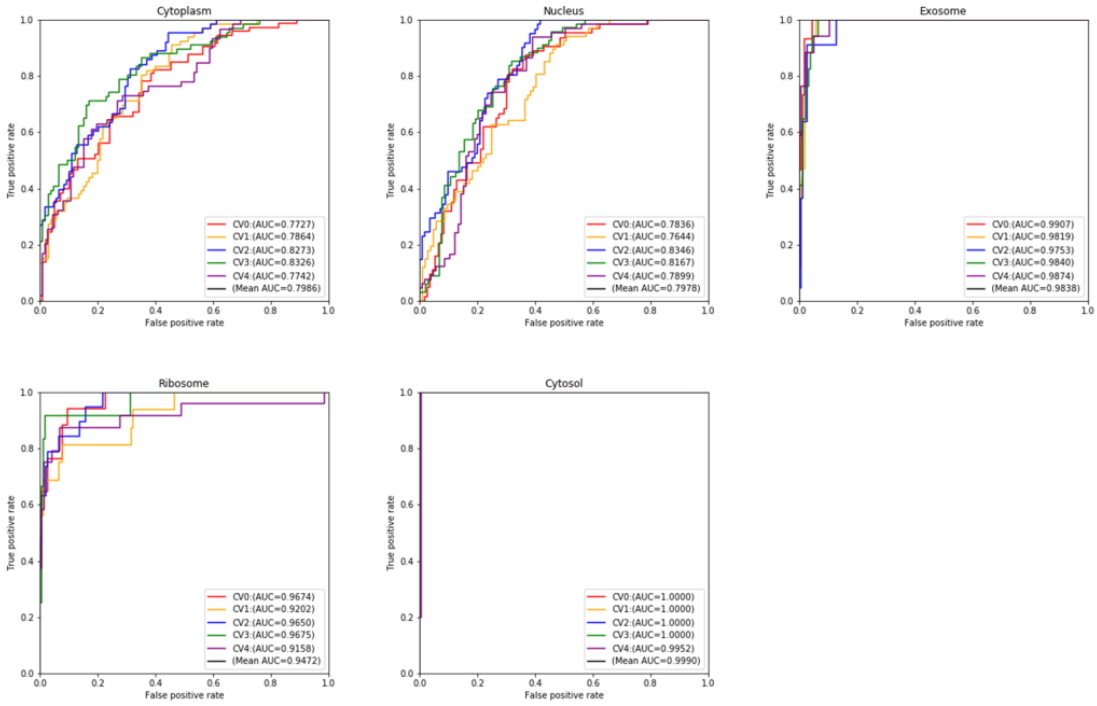

Figure 4. The 5-fold cross-validation ROC curve for each class with LightGBM-LncLoc.

Table 8. Performance for each class on the independent test.

Predictor	LightGBM-LncLoc			DeepLncLoc			iLoc-lncRNA		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Cytoplasm	0.667	0.400	0.500	0.778	0.350	0.483	0.553	0.700	0.618
Nucleus	0.429	0.900	0.581	0.400	0.800	0.533	0.467	0.350	0.400
Ribosome	0.800	0.400	0.533	0.500	0.400	0.444	0.333	0.500	0.316
Cytosol	1.000	0.500	0.667	0.833	0.500	0.625	null	null	null
Exosome	1.000	0.429	0.600	1.000	0.571	0.727	0.600	0.429	0.500

Note: F1 represents F-measure.

3.5. WebServer

We deployed LightGBM-LncLoc at the website http://www.biolscience.cn/LightGBM_LncLoc/ for the purpose of conveniently recognizing the subcellular localization of lncRNA. The only thing users need to do is submit the lncRNA sequences in FASTA format to the textbox or upload a fasta file, and then click the button "submit". It will take a certain time to return the predicted results. The consumed time will depend on the number of lncRNA sequences. In general, computing a sequence takes less than 5 s. The predictive

result is exhibited as a table that shows the probabilities of classifying the input into each subcellular localization type.

4. Conclusions

Identifying the subcellular localization of lncRNAs is critical to exploring their roles in cellular processes. Due to the limitations of current techniques and knowledge, it is a challenging task to precisely recognize the subcellular localization of lncRNAs. We developed a lightGBM-based computational predictor for recognizing lncRNA subcellular localization (LightGBM-LncLoc) and implemented it as a user-friendly webserver. The LightGBM-LncLoc obtained an average accuracy of 0.706 with five-fold cross-validation and an average of 0.567 with the independent test, outperforming two state-of-the-art methods: iLoc-lncRNA and DeepLncLoc. LightGBM-LncLoc is able to identify up to five categories of lncRNA subcellular localization.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math11030602/s1, Figure S1: Confusion matrix of LightGBM-Lncloc with DeepLncLoc on the test set. (a) LightGBM-Lncloc, (b) DeepLncLoc. In the confusion matrix, the horizontal coordinates represent the predicted labels and the vertical coordinates represent the true labels. The diagonal line represents the number of correctly predicted samples. Figure S2: Confusion matrix of LightGBM-Lncloc with iLoc-lncRNA on the test set. (a) LightGBM-Lncloc, (b) iLoc-lncRNA. In the confusion matrix, the horizontal coordinates represent the predicted labels and the vertical coordinates represent the true labels. The diagonal line represents the number of correctly predicted samples. Table S1. The performance evaluation results of 7 feature encoding using XGBoost; Table S2. The performance evaluation results of 7 feature encoding using SVM. Table S3. The performance evaluation results of 7 feature encoding using Random Forest. Table S4. The performance evaluation results of 7 feature encoding using Logical regression. Table S5. The performance evaluation results of 7 feature encoding using Multilayer perceptron. Table S6. The performance evaluation results of 5 optimal candidate base classifiers with their best feature combination. Table S7. Hyper-parameter values of the optimal candidate base classifiers. Table S8. The performance of 5 fold cross-validation under 10 different random seeds.

Author Contributions: J.L.: data curation, methodology, software, investigation, writing. P.Z.: methodology. Y.Q.: investigation. G.H.: conceptualization, funding acquisition, supervision, writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (62272310), by the Hunan Provincial Natural Science Foundation of China (2022JJ50177), by the Scientific Research Fund of Hunan Provincial Education Department (21A0466), and by Shaoyang University Innovation Foundation for Postgraduate (CX2022SY041).

Data Availability Statement: The data are available at https://github.com/rice1ee/LightGBM-LncLoc (accessed on 10 January 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Birney, E.; Stamatoyannopoulos, J.A.; Dutta, A.; Guigó, R.; Gingeras, T.R.; Margulies, E.H.; Weng, Z.; Snyder, M.; Dermitzakis, E.T.; Thurman, R.E.; et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **2007**, *447*, 799–816. [CrossRef]
2. Lu, C.; Yang, M.; Luo, F.; Wu, F.-X.; Li, M.; Pan, Y.; Li, Y.; Wang, J. Prediction of lncRNA–disease associations based on inductive matrix completion. *Bioinformatics* **2018**, *34*, 3357–3364. [CrossRef] [PubMed]
3. Kitagawa, M.; Kitagawa, K.; Kotake, Y.; Niida, H.; Ohhata, T. Cell cycle regulation by long non-coding RNAs. *Cell. Mol. Life Sci.* **2013**, *70*, 4785–4794. [CrossRef]
4. Brazão, T.F.; Johnson, J.S.; Müller, J.; Heger, A.; Ponting, C.P.; Tybulewicz, V.L. Long noncoding RNAs in B-cell development and activation. *Blood J. Am. Soc. Hematol.* **2016**, *128*, e10–e19. [CrossRef] [PubMed]
5. Delas, M.J.; Sabin, L.R.; Dolzhenko, E.; Knott, S.R.; Munera Maravilla, E.; Jackson, B.T.; Wild, S.A.; Kovacevic, T.; Stork, E.M.; Zhou, M.; et al. lncRNA requirements for mouse acute myeloid leukemia and normal differentiation. *eLife* **2017**, *6*, e25607. [CrossRef] [PubMed]

6. Sirey, T.M.; Roberts, K.; Haerty, W.; Bedoya-Reina, O.; Rogatti-Granados, S.; Tan, J.Y.; Li, N.; Heather, L.C.; Carter, R.N.; Cooper, S. The long non-coding RNA Cerox1 is a post transcriptional regulator of mitochondrial complex I catalytic activity. *eLife* **2019**, *8*, e45051. [CrossRef]
7. Sun, X.; Wong, D. Long non-coding RNA-mediated regulation of glucose homeostasis and diabetes. *Am. J. Cardiovasc. Dis.* **2016**, *6*, 17–25.
8. Statello, L.; Guo, C.-J.; Chen, L.-L.; Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **2021**, *22*, 159. [CrossRef]
9. Samarfard, S.; Ghorbani, A.; Karbanowicz, T.P.; Lim, Z.X.; Saedi, M.; Fariborzi, N.; McTaggart, A.R.; Izadpanah, K. Regulatory non-coding RNA: The core defense mechanism against plant pathogens. *J. Biotechnol.* **2022**, *359*, 82–94. [CrossRef]
10. Xing, C.; Sun, S.-g.; Yue, Z.-Q.; Bai, F. Role of lncRNA LUCAT1 in cancer. *Biomed. Pharmacother.* **2021**, *134*, 111158. [CrossRef]
11. Carlevaro-Fita, J.; Johnson, R. Global positioning system: Understanding long noncoding RNAs through subcellular localization. *Mol. Cell* **2019**, *73*, 869–883. [CrossRef] [PubMed]
12. Bridges, M.C.; Daulagala, A.C.; Kourtidis, A. LNCcation: lncRNA localization and function. *J. Cell Biol.* **2021**, *220*, e202009045. [CrossRef]
13. Kugel, J.F.; Goodrich, J.A. Non-coding RNAs: Key regulators of mammalian transcription. *Trends Biochem. Sci.* **2012**, *37*, 144–151. [CrossRef] [PubMed]
14. Melé, M.; Rinn, J.L. "Cat's Cradling" the 3D genome by the act of LncRNA transcription. *Mol. Cell* **2016**, *62*, 657–664. [CrossRef] [PubMed]
15. Saxena, A.; Carninci, P. Long non-coding RNA modifies chromatin: Epigenetic silencing by long non-coding RNAs. *Bioessays* **2011**, *33*, 830–839. [CrossRef]
16. Li, B.; Cai, L.; Liao, B.; Fu, X.; Bing, P.; Yang, J. Prediction of protein subcellular localization based on fusion of multi-view features. *Molecules* **2019**, *24*, 919. [CrossRef]
17. Alaa, A.; Eldeib, A.M.; Metwally, A.A. Protein Subcellular Localization Prediction Based on Internal Micro-similarities of Markov Chains. In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 1355–1358.
18. Gardy, J.L.; Brinkman, F.S. Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* **2006**, *4*, 741–751. [CrossRef]
19. Bhasin, M.; Garg, A.; Raghava, G.P.S. PSLpred: Prediction of subcellular localization of bacterial proteins. *Bioinformatics* **2005**, *21*, 2522–2524. [CrossRef]
20. Gardy, J.L.; Spencer, C.; Wang, K.; Ester, M.; Tusnady, G.E.; Simon, I.; Hua, S.; DeFays, K.; Lambert, C.; Nakai, K. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.* **2003**, *31*, 3613–3617. [CrossRef]
21. Hua, S.; Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **2001**, *17*, 721–728. [CrossRef]
22. Almagro Armenteros, J.J.; Sønderby, C.K.; Sønderby, S.K.; Nielsen, H.; Winther, O. DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics* **2017**, *33*, 3387–3395. [CrossRef]
23. Shen, Y.; Ding, Y.; Tang, J.; Zou, Q.; Guo, F. Critical evaluation of web-based prediction tools for human protein subcellular localization. *Brief. Bioinform.* **2020**, *21*, 1628–1640. [CrossRef] [PubMed]
24. Cao, Z.; Pan, X.; Yang, Y.; Huang, Y.; Shen, H.-B. The lncLocator: A subcellular localization predictor for long non-coding RNAs based on a stacked ensemble classifier. *Bioinformatics* **2018**, *34*, 2185–2194. [CrossRef] [PubMed]
25. Su, Z.-D.; Huang, Y.; Zhang, Z.-Y.; Zhao, Y.-W.; Wang, D.; Chen, W.; Chou, K.-C.; Lin, H. iLoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* **2018**, *34*, 4196–4204. [CrossRef] [PubMed]
26. Feng, S.; Liang, Y.; Du, W.; Lv, W.; Li, Y. LncLocation: Efficient subcellular location prediction of long non-coding RNA-based multi-source heterogeneous feature fusion. *Int. J. Mol. Sci.* **2020**, *21*, 7271. [CrossRef]
27. Gudenas, B.L.; Wang, L. Prediction of LncRNA subcellular localization with deep learning from sequence features. *Sci. Rep.* **2018**, *8*, 16385. [CrossRef]
28. Zeng, M.; Wu, Y.; Lu, C.; Zhang, F.; Wu, F.-X.; Li, M. DeepLncLoc: A deep learning framework for long non-coding RNA subcellular localization prediction based on subsequence embedding. *Brief. Bioinform.* **2022**, *23*, bbab360. [CrossRef]
29. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
30. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, UK, 2013; Volume 26.
31. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
32. Zhang, T.; Tan, P.; Wang, L.; Jin, N.; Li, Y.; Zhang, L.; Yang, H.; Hu, Z.; Zhang, L.; Hu, C. RNALocate: A resource for RNA subcellular localizations. *Nucleic Acids Res.* **2017**, *45*, D135–D138. [CrossRef]
33. Wen, X.; Gao, L.; Guo, X.; Li, X.; Huang, X.; Wang, Y.; Xu, H.; He, R.; Jia, C.; Liang, F. lncSLdb: A resource for long non-coding RNA subcellular localization. *Database* **2018**, *2018*, bay085. [CrossRef] [PubMed]

34. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef] [PubMed]
35. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
36. Chen, Z.; Zhao, P.; Li, F.; Marquez-Lago, T.T.; Leier, A.; Revote, J.; Zhu, Y.; Powell, D.R.; Akutsu, T.; Webb, G.I. iLearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinform.* **2020**, *21*, 1047–1057. [CrossRef] [PubMed]
37. Xu, H.; Jia, P.; Zhao, Z. Deep4mC: Systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning. *Brief. Bioinform.* **2021**, *22*, bbaa099. [CrossRef] [PubMed]
38. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, UK, 2017; Volume 30.
39. Wang, D.; Zhang, Y.; Zhao, Y. LightGBM: An effective miRNA classification method in breast cancer patients. In Proceedings of the 2017 International Conference on Computational Biology and Bioinformatics, Newark, NJ, USA, 18–20 October 2017; pp. 7–11.
40. Li, F.; Guo, X.; Jin, P.; Chen, J.; Xiang, D.; Song, J.; Coin, L.J.M. Porpoise: A new approach for accurate prediction of RNA pseudouridine sites. *Brief. Bioinform.* **2021**, *22*, bbab245. [CrossRef]
41. Emami, N.; Ferdousi, R. AptaNet as a deep learning approach for aptamer–protein interaction prediction. *Sci. Rep.* **2021**, *11*, 6074. [CrossRef]
42. Sperandei, S. Understanding logistic regression analysis. *Biochem. Med.* **2014**, *24*, 12–18. [CrossRef]
43. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
44. Tyree, S.; Weinberger, K.Q.; Agrawal, K.; Paykin, J. Parallel boosted regression trees for web search ranking. In Proceedings of the 20th International Conference on World Wide Web, Hyderabad, India, 28 March–1 April 2011; pp. 387–396.
45. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
46. Ridgeway, G. Generalized Boosted Models: A guide to the gbm package. *Update* **2007**, *1*, 2007.
47. Song, Y.; Jiao, X.; Qiao, Y.; Liu, X.; Qiang, Y.; Liu, Z.; Zhang, L. Prediction of double-high biochemical indicators based on LightGBM and XGBoost. In Proceedings of the 2019 International Conference on Artificial Intelligence and Computer Science, Wuhan, China, 12–13 July 2019; pp. 189–193.
48. Bi, Y.; Xiang, D.; Ge, Z.; Li, F.; Jia, C.; Song, J. An interpretable prediction model for identifying N7-methylguanosine sites based on XGBoost and SHAP. *Mol. Ther.-Nucleic Acids* **2020**, *22*, 362–372. [CrossRef] [PubMed]
49. Nguyen-Vo, T.-H.; Nguyen, Q.H.; Do, T.T.; Nguyen, T.-N.; Rahardja, S.; Nguyen, B.P. iPseU-NCP: Identifying RNA pseudouridine sites using random forest and NCP-encoded features. *BMC Genom.* **2019**, *20*, 971. [CrossRef] [PubMed]
50. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K. Xgboost: Extreme gradient boosting. *R Package Version 0.4-2* **2015**, *1*, 1–4.
51. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Appl.* **1998**, *13*, 18–28. [CrossRef]
52. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [CrossRef]
53. Wright, R.E. Logistic regression. In *Reading and Understanding Multivariate Statistics*; American Psychological Association: Washington, DC, USA, 1995; pp. 217–244.
54. Ruck, D.W.; Rogers, S.K.; Kabrisky, M. Feature selection using a multilayer perceptron. *J. Neural Netw. Comput.* **1990**, *2*, 40–48.