mathematics

Article

Prediction of Extracellular Matrix Proteins by Fusing Multiple Feature Information, Elastic Net, and Random Forest Algorithm

Minghui Wang [1,2], **Lingling Yue** [1,2], **Xiaowen Cui** [1,2], **Cheng Chen** [1,2], **Hongyan Zhou** [1,2], **Qin Ma** [3] **and Bin Yu** [1,2,4,*]

[1] College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China; mhwang@yeah.net (M.W.); llyue_qust@126.com (L.Y.); xwcui_qust@126.com (X.C.); chenchengqust@163.com (C.C.); zhoughongyan@qust.edu.cn (H.Z.)

[2] Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China

[3] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA; qin.ma@osumc.edu

[4] School of Life Sciences, University of Science and Technology of China, Hefei 230027, China

* Correspondence: yubin@qust.edu.cn

Received: 3 January 2020; Accepted: 23 January 2020; Published: 31 January 2020

Abstract: Extracellular matrix (ECM) proteins play an important role in a series of biological processes of cells. The study of ECM proteins is helpful to further comprehend their biological functions. We propose ECMP-RF (extracellular matrix proteins prediction by random forest) to predict ECM proteins. Firstly, the features of the protein sequence are extracted by combining encoding based on grouped weight, pseudo amino-acid composition, pseudo position-specific scoring matrix, a local descriptor, and an autocorrelation descriptor. Secondly, the synthetic minority oversampling technique (SMOTE) algorithm is employed to process the class imbalance data, and the elastic net (EN) is used to reduce the dimension of the feature vectors. Finally, the random forest (RF) classifier is used to predict the ECM proteins. Leave-one-out cross-validation shows that the balanced accuracy of the training and testing datasets is 97.3% and 97.9%, respectively. Compared with other state-of-the-art methods, ECMP-RF is significantly better than other predictors.

Keywords: extracellular matrix protein; multi-information fusion; synthetic minority oversampling technique; elastic net; random forest

1. Introduction

The extracellular matrix is a macromolecule synthesized by animal cells that is distributed on the cell surface or between cells [1,2]. The ECM not only provides structural support to cells within the tumor but also provides anchorage and tissue separation for the cells. It also has a coherence effect to mediate communication between cells, and it contributes to survival and differentiation signals [3,4]. ECM proteins actively promote essential cellular processes such as differentiation, proliferation, adhesion, migration, and apoptosis [5–8]. ECM proteins in certain parts of etiology are the drivers of disease [9]. Defects of ECM proteins are associated with many human diseases (such as cancer, atherosclerosis, asthma, fibrosis, and arthritis), and modified ECM proteins help to understand complex pathologies [10,11]. ECM protein research contributes to the development of novel cell-adhesive biomaterials that are critical in many medical fields, such as cell therapy or tissue engineering [12,13]. The diversity of ECM proteins contributes to the diversity of ECM function; thus, they play a comprehensive biological role in the basic life activities of cells. Therefore, accurate recognition of ECM proteins plays a crucial role in physiology and pathology research.

ECM protein prediction is an important branch in the field of protein function research. Compared with time-consuming and laborious experimental methods, bioinformatics methods can integrate advanced machine learning methods to accurately and efficiently predict proteins by fusing sequence information, physicochemical properties, structural information, and evolutionary information. The construction of protein prediction models mainly focuses on three aspects. (1) Feature extraction of protein sequences. The amino-acid letter sequence is converted into a numerical sequence by a feature extraction method. At present, protein sequence feature extraction methods are mainly based on sequence information, physicochemical properties, evolutionary information, and structural information. (2) Feature selection. Eliminating redundant features by feature selection technology can not only reduce the dimension of the feature space but also improve the robustness of the prediction model. More importantly, in predictive models, feature selection can assess the importance of features and help discover the molecular mechanisms via which proteins perform their corresponding biological functions. (3) Model construction and model evaluation. A prediction algorithm is designed and a prediction model is built. An effective prediction algorithm can improve the generalization ability and prediction performance of the model. Because the machine learning method has the ability to train the model and prediction, it is widely used in proteomics. Current mainstream machine learning methods include random forest (RF) [14], naïve Bayes [15,16], decision tree (DT) [17], support vector machine (SVM) [18], extreme gradient boosting (XGBoost) [19], adaptive boosting (AdaBoost) [20], logistic regression (LR) [21], gradient boosting decision tree (GBDT) [22], etc.

Given the important research significance of ECM proteins in physiopathology, more and more researchers are committed to the prediction research of ECM proteins. Table 1 summarizes the prediction methods of ECM proteins.

Table 1. The existing extracellular matrix (ECM) proteins prediction methods.

Author(s)	Method	Performance Parameters
Jung et al. [23]	91 features that were conventionally used by localization prediction methods and five new features are introduced to describe the unique features of ECM proteins.	receiver operating characteristic (ROC)
Anitha et al. [24]	Anitha [24] used position-specific scoring matrix (PSSM) to extract protein sequence features and used the support vector machine (SVM^{hmm}) classifier.	sensitivity (Sn), specificity (Sp), matthew's correlation coefficient (MCC), accuracy (ACC)
Kandaswamy et al. [25]	ECM proteins were recognized from sequence-derived properties.	ROC, Sn, Sp, MCC, ACC
Zhang et al. [26]	This method explored a variety of sequence-based discriminative features including evolutionary information, structural information, and physicochemical properties, then used Fisher-Markov selectors and incremental feature selection methods to search for the optimal feature vector, and finally built a model through SVM.	Sn, Sp, MCC, ACC
Yang et al. [27]	A hybrid feature integration method based on random forest was proposed.	ACC, Sn, Sp, balanced accuracy (BACC)
Ali et al. [28]	k-nearest neighbor (KNN) was combined with the hybrid feature space of pseudo amino-acid composition (PseAAC) and dipeptide composition (DPC).	ACC, Sn, Sp, precision, recall, F-measure
Kabir et al. [29]	This method extracted sequence evolution information based on a gray system model.	ACC, Sn, Sp, BACC

At present, a series of research results were obtained by predicting ECM proteins through bioinformatics and machine learning methods, but there is still room for improvement. Firstly, the development of an effective feature extraction method can achieve higher prediction accuracy. Extracting accurate feature vectors for proteins is a key step in successful prediction. Secondly, how to delete redundant information caused by feature fusion while retaining valid feature information is also one of the important directions of our research. The feature selection methods used in this paper include principal component analysis (PCA) [30], factor analysis (FA) [31], mutual information (MI) [32], least absolute shrinkage and selection operator (LASSO) [33], kernel principal component analysis (KPCA) [34], logistic regression (LR) [35], and elastic net (EN) [36,37]. Finally, how to choose prediction methods and tools with high calculation efficiency and good performance is also the focus of our research. Different from other methods, we explore a variety of sequence-based discriminant features, including sequence information, evolutionary information, and physicochemical properties. We use the synthetic minority oversampling technique (SMOTE) to deal with class imbalance between data. Instead of simply combining features that may cause information redundancy and unwanted noise, the optimal feature subset is selected by EN. We choose an RF classifier which hinders over-fitting in the prediction of ECM proteins.

In this paper, a new ECM proteins prediction method named ECMP-RF (extracellular matrix proteins prediction by random forest) is proposed. Firstly, we integrate encoding based on grouped weight (EBGW), pseudo amino-acid composition (PseAAC), pseudo position-specific scoring matrix (PsePSSM), a local descriptor (LD), and an autocorrelation descriptor (AD) to extract protein sequence features. Five kinds of feature coding information are fused as the initial feature space. Then, the SMOTE method is used to balance the samples. Thirdly, comparing seven different feature selection methods for prediction, the EN is used to select the feature vectors. Finally, the RF classifier is used to predict ECM proteins. Based on the prediction results, RF is selected as a classifier to construct a multi-feature fusion ECM protein prediction model. The balanced accuracy (BACC) obtained by leave-one-out cross-validation (LOOCV) on the training dataset and the testing dataset is as high as 97.28% and 97.94%, respectively. The prediction effect is higher than the currently known classification models. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/ECMP-RF/.

2. Materials and Methods

2.1. Datasets

A dataset is mainly used for training and testing models. The quality of datasets has a direct effect on model construction. Therefore, establishing an ECM protein prediction model requires an objective and representative dataset. To compare with the existing research results, this paper uses a dataset constructed in the existing literature for experimental analysis, thereby ensuring the scientific relevance and fairness of the comparative analysis. Yang [27] took the dataset established by Kandaswamy [25] as the original training dataset and constructed a training dataset and testing dataset according to the principle of dataset construction. The training dataset and testing dataset contain 410 ECM proteins and 85 ECM proteins, respectively. At the same time, the training dataset and testing dataset contain 4464 non-ECM proteins and 130 non-ECM proteins, respectively.

2.2. Encoding Based on Grouped Weight

Zhang [38] proposed the encoding based on grouped weight (EBGW) method to extract the characteristics of the physicochemical properties of proteins. EBGW is widely used in research fields such as protein interactions [39,40], prediction of Golgi protein types [41], and prediction of protein mitochondrial localization [42]. EBGW is based on the physicochemical properties of amino-acid sequences. It uses the concept of "coarse graining" to convert protein sequences into binary characteristic sequences. The group weight coding of protein sequences is realized by introducing a regular weight function. Amino acids are the basic units of protein. The protein sequence is composed of 20 amino acids through a series of biochemical reactions. Different amino acids have different physicochemical properties. According to different physicochemical properties, amino acids can be divided into four categories. These are acidic amino acids $C_1 = \{\mathrm{D,E}\}$, basic amino acids $C_2 = \{\mathrm{H,K,R}\}$, neutral and polar amino acids $C_3 = \{\mathrm{Q,N,S,T,Y,C}\}$, and neutral and hydrophobic amino acids $C_4 = \{\mathrm{A,F,G,I,L,M,P,V,W}\}$. The above four types of amino acids can be combined one by one to obtain three new division methods: $\{C_1, C_3\}$ and $\{C_2, C_4\}$, $\{C_1, C_2\}$ and $\{C_3, C_4\}$, $\{C_1, C_4\}$ and $\{C_2, C_3\}$. After a protein is mapped according to these three division methods, the protein sequence is reduced to a binary feature sequence.

The protein sequence $\alpha = a_1 a_2 \cdots a_n$ is transformed into three binary sequences according to Equations (1)–(4).

$$\varphi_i(\alpha) = \varphi_i(a_1)\varphi_i(a_2)\cdots\varphi_i(a_n), \tag{1}$$

$$\varphi_1(a_j) = \begin{cases} 1 & if \quad a_j \in \{C_1, C_2\} \\ 0 & if \quad a_j \in \{C_3, C_4\} \end{cases}, \tag{2}$$

$$\varphi_2(a_j) = \begin{cases} 1 & if \quad a_j \in \{C_1, C_3\} \\ 0 & if \quad a_j \in \{C_2, C_4\} \end{cases}, \tag{3}$$

$$\varphi_3(a_j) = \begin{cases} 1 & if \quad a_j \in \{C_1, C_4\} \\ 0 & if \quad a_j \in \{C_2, C_3\} \end{cases}, \tag{4}$$

where $\varphi_i(\alpha)$ represents protein sequence characteristics. Given the sequence feature $\varphi_i(\alpha)$ of length n, weight refers to the number of occurrences of "1" in the sequence, and rule weight refers to the frequency of occurrence. By dividing the raw sequence into L subsequences, these features can be obtained. Then, the normal weight of each subsequence is calculated, thereby obtaining an L-dimensional vector H_i. The transformation from $\varphi_i(\alpha)$ to H_i is the group weight coding of sequence features. Thus, the three sequence features are used to obtain three vectors, and the three vectors are combined to obtain a $3L$-dimensional vector, denoted as $W = [w_1, w_2, \cdots, w_{3L}]$, where W is referred to as the group weight encoding of protein sequence α.

2.3. Pseudo Amino-Acid Composition

Chou [43] proposed the pseudo amino-acid composition (PseAAC) to solve the problem of feature vector extraction in protein sequences. At present, PseAAC is widely used by researchers in proteomic fields, such as protein site prediction [44], protein structure prediction [41,45,46], protein subcellular localization prediction [47–50], and protein submitochondrion prediction [42,51]. The specific calculation steps of the PseAAC algorithm are shown in Equations (5) and (6).

$$P = [P_1, P_2, \cdots, P_{20}, P_{20+1}, \cdots, P_{20+\lambda}]^T, \tag{5}$$

where,

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{u=1}^{20} f_u + \omega \sum_{k=1}^{\lambda} \tau_k} & 1 \le u \le 20 \\ \dfrac{\omega \tau_{u-20}}{\sum_{u=1}^{20} f_u + \omega \sum_{k=1}^{\lambda} \tau_k} & 20+1 \le u \le 20+\lambda \end{cases}, \tag{6}$$

where P represents the eigenvector, τ_k refers to the nearest correlation factor with a number of k, and ω refers to the weight factor with a value of 0.05 [52]. f_u represents the frequency of the μ-th amino acid in this sequence $(u = 1, 2, \cdots, 20)$. In this method, the protein sequence is mapped to an ($20+\lambda$)-dimensional vector; the first 20 dimensions are the AAC, while the latter λ-dimensional is the sequence correlation factor, which can be obtained through the physicochemical properties of amino acids. The λ value cannot be greater than the shortest length of the protein sequence in the training set; thus, the λ value is between 10 and 50.

2.4. Pseudo Position-Specific Scoring Matrix

The pseudo position-specific scoring matrix (PsePSSM) was proposed by Chou and Shen. [53]. PsePSSM can not only extract sequence information, but also reflect sequence evolution information. It is extensively used in proteomics research. Qiu [54] predicted the location of protein mitochondria by integrating PsePSSM into general PseAAC. Shi [55] used PsePSSM to extract the features of a drug target. The method of PsePSSM is used to extract the features of a protein sequence. We use the multi-sequence homology matching tool position-specific iterated basic local alignment search tool (PSI-BLAST) to set the parameters to three iterations, and the PSSM of each protein sequence is shown in Equation (7).

$$P_{\mathrm{PSSM}} = \begin{pmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{i,1} & P_{i,2} & \cdots & P_{i,20} \\ \vdots & \vdots & \vdots & \vdots \\ P_{L,1}, & P_{L,2} & \cdots & P_{L,20} \end{pmatrix}. \tag{7}$$

The elements in a PSSM matrix are normalized by Equation (8).

$$f(x) = 1/\left(1+e^{-x}\right). \tag{8}$$

Because the length L of the protein sequence is different, the PSSM matrix with different length is transformed into a matrix with the same dimension using Equations (9) and (10), and then the PsePSSM is calculated.

$$p^{m}_{\mathrm{PsePSSM}} = (\theta^{m}_{1}, \theta^{m}_{2}, \cdots, \theta^{m}_{j}, \cdots \theta^{m}_{20})^{T}, \tag{9}$$

$$\theta^{m}_{j} = \frac{1}{L-m}\sum_{i=1}^{L-m}(p_{i,j} - p_{(i+m),j})^{2} \quad (j = 1, 2, \cdots 20; m < L, m \neq 0), \tag{10}$$

$$P_{\mathrm{PsePSSM}} = (\overline{p_1}, \overline{p_2}, \cdots, \overline{p_{20}}, \theta^{1}_{1}, \theta^{1}_{2}, \cdots, \theta^{1}_{20}, \cdots, \theta^{m}_{1}, \theta^{m}_{2}, \cdots, \theta^{m}_{20})^{T}. \tag{11}$$

PsePSSM generates a $20+20\times m$-dimensional feature vector from a protein sequence, which can change the length of different protein sequences in the feature extraction data set into a vector with a uniform dimension. Restricted by the shortest length of the protein sequence, m values range from 10 to 49.

2.5. Local Descriptor

The local descriptor (LD) algorithm considers the discontinuity of the sequence [37]. The LD algorithm represents the effective feature information of a protein sequence through amino-acid classification, protein sequence division, protein folding, and other methods, so as to provide important sequence information and physical and chemical property feature information for ECM proteins prediction.

Each protein sequence is divided into 10 segments. For each specific local segment, composition (C), transition (T), and distribution (D) are calculated. C refers to the frequency of amino acids with specific properties in 10 segments. T refers to the frequency of dipeptides composed of seven groups of amino acids in 10 segments. D refers to the distribution pattern of the first, 25%, 50%, 75%, and last amino acids in each protein sequence. The amino-acid groups are shown in Table 2.

For each region, C, T, and D are computed, and 63 features are obtained, among which $C=7$, $T=21$, and $D=35$. Then, 10 local segments generate $10\times 63=630$ features; thus, the LD algorithm constructs a 630-dimensional vector.

Table 2. Seven groups of amino acids.

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
C	R, K	E, D	A, G, V	H, N, W, Q	P, L, I, F	T, S, M, Y

2.6. Autocorrelation Descriptor

By using the physicochemical properties of amino acids, protein sequences can be numerical, but the length of the numerical sequences obtained is not the same. The autocorrelation descriptor (AD) [56,57] can transform numerical sequences with different lengths into feature vectors with the same length using an appropriate coding method, so as to extract the sequence information of proteins effectively and improve the prediction accuracy. The three descriptors used in this study are described below.

(1) Definition of Moreau-Broto autocorrelation descriptor.

$$\mathrm{NMBA}(l)=\frac{\mathrm{NBA}(l)}{N-l}\quad l=1,2,\cdots,lag, \tag{12}$$

where $\mathrm{MBA}(l)=\sum_{i=1}^{N-l}P(AA_i)P(AA_{i+l})$, AA_i and AA_{i+l} represent the amino acids of i and $i+l$ of the protein sequence, respectively, $P(AA_i)$ and $P(AA_{i+l})$ represent the standardized physicochemical values of i and $i+l$ of AA_i and AA_{i+l}, respectively, and lag is the parameter to be adjusted, which represents the length of sliding window.

(2) Definition of Moran autocorrelation descriptor.

$$\mathrm{MA}(l)=\frac{\frac{1}{N-l}\sum_{i=1}^{N-l}(P(AA_i)-\tilde{P})(P(AA_{i+l})-\tilde{P})}{\frac{1}{N}\sum_{i=1}^{N}(P(AA_i)-\tilde{P})^2}\quad l=1,2,\cdots,lag, \tag{13}$$

where $\tilde{p}$ represents the average value of the physicochemical properties of the whole protein sequence.

(3) Definition of the Geary autocorrelation descriptor.

$$\mathrm{GA}(l)=\frac{\frac{1}{2(N-l)}\sum_{i=1}^{N-l}(P(AA_i)-P(AA_{i+l}))^2}{\frac{1}{N}\sum_{i=1}^{N}(P(AA_i)-\tilde{P})^2}\quad l=1,2,\cdots,lag. \tag{14}$$

Using the above three autocorrelation descriptors, we can extract a $3\times n\times lag$ -dimensional feature vector (n is the number of amino acid physicochemical properties used in this paper). The built-in parameters lag are 10, 20, 30, 40, and 50.

2.7. Synthetic Minority Oversampling Technique

The imbalance of categories has a negative impact on the classification performance of the model. A small number of samples provides less "information", which leads to a low prediction performance of a small number of samples; thus, the prediction results tend to the majority. There are significant sample imbalances in the selected datasets of this study. Among them, there are 11-fold more non-ECM proteins in the training dataset than ECM proteins, and there are 1.5-fold more non-ECM proteins in the testing dataset than ECM proteins, which results in poor prediction accuracy for a few categories. In order to solve the problem of sample imbalance and make full use of the sample information in the original datasets, Chawla [58] proposed the synthetic minority oversampling technique (SMOTE) algorithm. This is a common oversampling method [59], which is based on "interpolation" to synthesize new samples for a few samples and add them to the dataset, so as to achieve the balance of dataset samples and effectively solve the problem of classification overfitting.

The algorithm flow of SMOTE is described below. If the number of samples of a minority class in the training dataset is set as T, then SMOTE synthesizes NT (N is a positive integer) new samples for this minority class. If $N<1$ is given, the algorithm forces $N=1$ and "considers" the sample number of a few classes as T. Consider a sample i of the minority class with characteristic vectors of $x_i, i \in \{1,\cdots,T\}$; then, the following steps are applied:

(1) Find k nearest neighbors of sample x_i from all T samples of the minority class (e.g., using Euclidean distance), and record them as $x_{i(near)}, near \in \{1,\cdots,k\}$.

(2) From k nearest neighbors, select a sample $x_{i(nn)}$ randomly, and regenerate it into a random number ξ_1 between 0 and 1, so as to synthesize a new sample x_{i1}: $x_{i1} = x_i + \xi_1 \times (x_{i(nn)} - x_i)$.

(3) Repeat step (2) N times, and synthesize N new samples $x_{inew}, new \in \{1,\cdots,N\}$. The above operations are carried out on all T minority samples, and NT new samples are obtained.

At present, SMOTE is widely used in bioinformatics research, for example, protein-protein interaction site prediction [39,55], protein category prediction [41], predicting protein submitochondrial localization [42], etc.

2.8. Elastic Net

The high-dimensional data fused by the feature extraction method contain a lot of redundant information. The elastic net (EN) algorithm eliminates redundant information, and it transforms the high-dimensional data into low-dimensional data while retaining the effective information of the original data. EN [36,37] is a linear regression model trained with L_1, L_2 norm as a prior regular term.

For the dataset $D = \{(x_1,y_1),(x_2,y_2),\cdots(x_m,y_m)\}$, taking the square error as the loss function, the optimization objective is shown in Equation (15).

$$\min_{\omega} \sum_{i=1}^{m} (y_i - \omega^T x_i)^2 . \tag{15}$$

However, when there are many sample features, a small number of samples causes overfitting; thus, it can be regularized. $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \dfrac{\lambda}{\lambda_1 + \lambda_2}$ are obtained by using a convex linear combination of L_1 norm and L_2 norm. In this case, Equation (15) becomes the following loss function:

Loss function:

$$\min_{\omega} \sum_{i=1}^{m} (y_i - \omega^T x_i)^2 + \lambda\alpha \|\omega\|_1 + (1-\alpha)\|\omega\|_2^2 . \tag{16}$$

2.9. Random Forest

As a representative machine learning method, the random forest (RF) algorithm is widely used in the field of bioinformatics [39,55]. It combines the theory of bagging and random subspace, integrates many decision trees to predict, and then votes the predicted values of each decision tree to get the final prediction results. Therefore, RF has strong classification ability. The specific algorithm flow of RF is given below.

(1) The original dataset is N. Bootstrapping is used to construct a classification tree. The remaining samples after random selection constitute the out-of-bag data.

(2) At first, m_{all} variables are randomly selected from each node of each tree. Then, a constant m_{try} ($m_{try} \ll m_{all}$) is set, and m_{try} variables are randomly selected from m_{all} variables. When the tree is split, the variable with the most classification ability is selected from m_{try} variables according

to the Gini criterion of node impurity measurement. The threshold value of variable classification is determined by checking each classification point.

(3) Every tree grows to its maximum without any pruning.

(4) An RF is made up of many classification trees. The outcome of classification is resolved by the voting result of the tree classifier.

Two important parameters of RF are the number of building classification trees $ntree$ and the size of the leaf node $nodesize$. In this study, $ntree = 150$ and $nodesize = 10$ were set.

2.10. Performance Evaluation

In statistical theory, LOOCV, independence test, and k-fold cross-validation are used to assess models. Among them, LOOCV can return the unique results of a given benchmark dataset, and it is widely used in proteomics research due to its strict response algorithm's accuracy and generalization ability [41,42]. Due to the imbalance of datasets, the result of the classifier tends to most classes, which leads to the high specificity (Sp) of the model. However, a good prediction system should have both high sensitivity (Sn) and high Sp; thus, the accuracy is not a very suitable evaluation index. In addition to accuracy (ACC), Sn, and Sp, this paper introduces BACC as the main performance index. ACC is the proportion of all correctly predicted samples in the total samples. Sn represents the ability of the model to predict positive samples, and Sp represents the ability of the model to predict negative samples, while BACC can comprehensively consider Sn and Sp.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}}, \tag{17}$$

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{18}$$

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{19}$$

$$\mathrm{BACC} = \frac{1}{2}(\mathrm{Sn} + \mathrm{Sp}). \tag{20}$$

In addition, the receiver operating characteristic (ROC) is a curve based on Sn and Sp, and the area under the curve (AUC) is the area under the ROC curve. A more robust prediction performance of the model achieves an AUC value closer to 1. The area under precision recall (AUPR) is also an important indicator for evaluating models, representing the area under the precision recall (PR) curve. The PR curve can intuitively show the recall and precision of the classifier on the entire sample. When one classifier's PR curve can wrap another classifier's PR curve, the former has better classification performance.

Figure 1 shows the calculation flow of ECMP-RF. The experimental environment was as follows: Windows Server 2012R2 Intel (R) Xeon (TM) central processing unit (CPU) E5-2650 @ 2.30 GHz 2.30 GHz with 32.0 GB of random-access memory (RAM), with Python3.6 and MATLAB2014a programming implementation.

The prediction steps of ECMP-RF are as follows.

(1) Obtain the datasets, and then input the protein sequences of different classes and the corresponding labels.

(2) Feature extraction. Treat protein sequences as special strings, and then convert character signals into numerical signals by encoding as follows: (a) extraction of protein sequence features using PseAAC. (b) feature extraction of protein evolutionary information using PsePSSM. (c) extraction of physicochemical properties of amino acids by EBGW. (d) extraction of physicochemical properties of proteins using AD. (e) extraction of protein sequence information and physicochemical properties using LD. The five extracted features are then fused, and a 1950-dimensional vector space is constructed for each protein sequence in the dataset.

(3) The problem of imbalance in datasets is solved by SMOTE.

(4) Feature selection. The EN is used to remove the redundancy and noise generated by feature fusion, and to screen the optimal feature vector, so as to provide good feature information.

(5) The selected subset of the best features is input into the RF classifier to predict ECM proteins.

(6) Model performance evaluation. ACC, Sn, SP, and BACC are used as evaluation indexes, and the prediction performance of the model is tested by LOOCV.

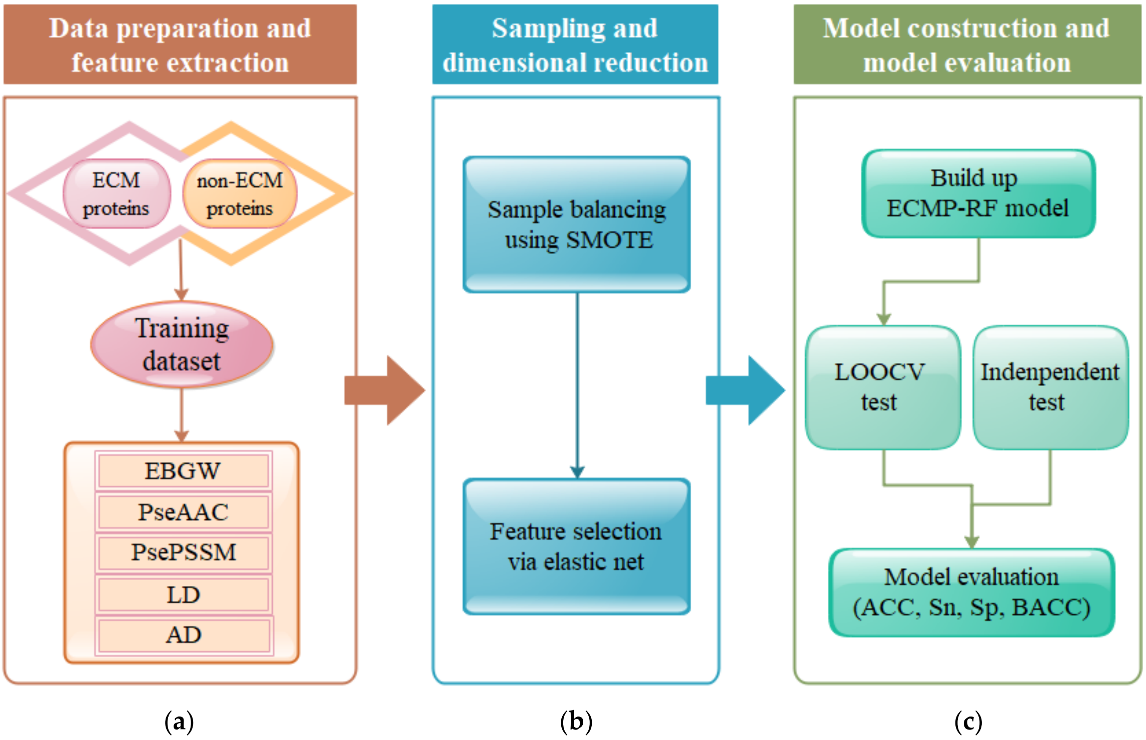
Data preparation and feature extraction
ECM proteins
non-ECM proteins
Training dataset
EBGW
PseAAC
PsePSSM
LD
AD
Sampling and dimensional reduction
Sample balancing using SMOTE
Feature selection via elastic net
Model construction and model evaluation
Build up ECMP-RF model
LOOCV test
Indenpendent test
Model evaluation (ACC, Sn, Sp, BACC)
(a)
(b)
(c)

Figure 1. The flowchart of the ECMP-RF method: (**a**) data preparation and feature extraction; (**b**) sampling and dimensional reduction; (**c**) model construction and model evaluation.

3. Results

3.1. Selection of Optimal Parameters λ, m, and lag

The selection of parameters is the core of the prediction model. Extracting effective feature information from protein sequences is a key step in constructing an ECM protein prediction model. In order to obtain better feature information, the parameters need to be continuously adjusted. In this paper, the training dataset was used as the research object to perform feature extraction on protein sequences. The selection of parameters λ in the PseAAC, m in the PsePSSM, and lag in the AD play an important role in model construction. If λ, m, and lag are selected too small, the extracted sequence information is reduced, and it cannot effectively describe the characteristics of protein sequence. If the values of parameters λ, m, and lag are too large, it causes noise interference and affects the prediction results. In order to find the best parameters, the λ value was set to 10, 20, 30, 40, and 50, the m value was set to 10, 20, 30, 40, and 49, and the lag value was set to 10, 20, 30, 40, and 50. The RF classifier was used to classify the dataset. ACC, Sn, Sp, and BACC were used as the evaluation indicators. The results were tested by the method of LOOCV. The evaluation values obtained by selecting different parameters in the PseAAC algorithm, PsePSSM algorithm, and AD algorithm are shown in Tables 3–5, respectively. The change in BACC value obtained by the feature extraction algorithm upon selecting different parameters is shown in Figure S1 (Supplementary Materials).

The λ value in PseAAC encoding reflects sequence correlation factors at different levels of amino-acid sequence information. Different values of λ have different effects on the prediction performance of the model. Considering that the sample sequence length was 50, the λ values in PseAAC encoding were set to 10, 20, 30, 40, and 50. As can be seen from Table 3, the evaluation indicators changed with λ. ACC, Sp, and BACC reached the highest when the value of λ was 20. At this time, the value of ACC was 92.63%, the value of Sn was 12.71%, and the value of BACC was 56.33%. In summary, λ=20 was selected as the best parameter of PseAAC coding. At this time, the performance of the model was best. When λ was 20, each protein sequence generated a 40-dimensional feature vector.

Different values of the parameter m of the PsePSSM encoding have a certain impact on the prediction performance of the model. Considering that the minimum sample sequence length was 50, the value of parameter m in PsePSSM encoding was set to 10, 20, 30, 40, and 49 in this order. It can be seen from Table 4 that the evaluation index changed with the value of m. When the value of m in the training dataset was 30, the maximum value of BACC was 60.44%, which means that PsePSSM had the best influence on the model performance. Therefore, the value of m = 30 was the best parameter in PsePSSM. At this time, each protein sequence generated a 620-dimensional feature vector.

Different lag values in AD have different effects on the prediction performance of the model. The lag values in AD were set to 10, 20, 30, 40 and 50 in turn. Table 5 shows that ACC, Sn, Sp, and BACC changed with the change in lag. When lag was 20 and 40, the values of ACC, Sp, and BACC in the training dataset slightly decreased, and when lag was 30, BACC was the highest, reaching 57.63%. When lag was 30, each protein sequence generated a 630-dimensional feature vector.

Table 3. The prediction overall accuracy of ECM proteins by selecting different λ values.

Performance Evaluation	λ = 10	λ = 20	λ = 30	λ = 40	λ = 50
ACC	92.58	92.63	92.61	92.53	92.55
Sn	11.74	12.71	12.22	11.24	11.73
Sp	100.00	99.96	99.98	99.98	99.96
BACC	55.87	56.33	56.10	55.61	55.85

Table 4. The prediction overall accuracy of ECM proteins by selecting different m values.

Performance Evaluation	m = 10	m = 20	m = 30	m = 40	m = 49
ACC	92.41	92.49	93.02	94.49	92.49
Sn	11.98	12.71	21.27	13.45	12.71
Sp	99.78	99.80	99.60	99.73	99.80
BACC	55.88	56.26	60.44	56.59	56.26

Table 5. The prediction overall accuracy of ECM proteins by selecting different lag values.

Performance Evaluation	lag = 10	lag = 20	lag = 30	lag = 40	lag = 50
ACC	92.74	92.69	92.78	92.76	92.76
Sn	15.16	14.67	15.40	14.67	15.40
Sp	99.84	99.84	99.87	99.91	99.84
BACC	57.50	57.26	57.63	57.29	57.62

3.2. The Effect of Feature Extraction Algorithm on Prediction Results

Extracting protein sequence information using the feature extraction algorithm is an important stage in the construction of ECM protein prediction models. In this paper, five feature extraction algorithms were selected: PseAAC, PsePSSM, EBGW, AD, and LD. In order to better learn the performance of the model, this article made comparisons using five separate feature coding methods and a hybrid feature coding method named ALL (EBGW + PseAAC + PsePSSM + LD + AD). The feature vectors extracted by the six feature extraction methods were input into the RF classifier to predict them respectively. ACC, BACC, Sn, and Sp were used as evaluation indicators to assess the performance of ECMP-RF. LOOCV was used for evaluation. The results predicted by different feature extraction methods are shown in Table 6.

In the training dataset, the ACC maximum and BACC maximum were 93.62% and 62.76% respectively, which were obtained by the LD algorithm. The ALL feature extraction method obtained an ACC of 95.03%, which was 1.41% to 2.60% higher than the result obtained by a single feature prediction. The BACC was 71.53%, which is 8.77%–15.61% higher than the result obtained by a single feature prediction. The difference in prediction results reflects the effect of feature extraction algorithms on the robustness of the prediction model. The ALL feature extraction method can not only express protein evolutionary information but also effectively use protein physicochemical properties, sequence information, and structure information. Compared with the single feature extraction method, it has better prediction performance.

Table 6. Prediction results of different characteristics of the training dataset.

Feature	ACC (%)	Sn (%)	Sp (%)	BACC (%)
EBGW	92.43	11.98	99.80	55.89
PseAAC	92.63	12.71	99.96	56.34
PsePSSM	93.02	21.27	99.60	60.44
LD	93.62	25.67	99.84	62.76
AD	92.78	15.40	99.87	57.64
ALL	95.03	43.28	99.78	71.53

3.3. Influence of Class Imbalance on Prediction Results

The two datasets used in this article, i.e., the training dataset and testing dataset, were highly imbalanced, with the ratio of positive and negative samples reaching 1:11 and 1:1.5 respectively. In this case, there was a serious imbalance between the dataset categories, and the prediction results of the classifier would be biased toward a large number of categories, which would adversely affect the prediction performance of the model. Therefore, we used SMOTE to improve the generalization ability of model. SMOTE achieves sample balance by deleting or adding some samples. LOOCV and RF classification algorithms were used to test balanced and unbalanced data. The prediction results are shown in Table 7.

Table 7. The impact of synthetic minority oversampling technique (SMOTE) on the prediction results of the training dataset.

Method	ACC (%)	Sn (%)	Sp (%)	BACC (%)
ALL	95.03	43.28	99.78	71.53
ALL (SMOTE)	97.04	98.75	95.33	97.04

Because of the imbalance of datasets, the evaluation of ACC index was biased. At the same time, the number of non-ECM proteins was much larger than the number of ECM proteins; thus, Sn and SP indexes could not accurately express the performance of the ECMP-RF, while the BACC could reasonably measure the performance of the model in all the above indicators. It can be seen in Table 7 that the dataset processed by SMOTE significantly improved the BACC. On the imbalanced training dataset, the BACC was 71.53%. After the dataset was balanced, the BACC was 97.04%, which is 25.51% higher than before. As a result, we think that, after SMOTE processing, the prediction performance of ECMP-RF was greatly improved.

3.4. The Effect of Feature Selection Algorithm on Prediction Results

Protein sequence features were extracted by EBGW, PseAAC, PsePSSM, LD, and AD feature extraction methods, and a 1950-dimensional feature vector was obtained after fusion. Since the fused feature vector had a large dimension and redundancy, it was very important to eliminate redundant information and keep the effective information of original data. Choosing different feature selection methods had a certain impact on the ACC of ECM protein prediction. To this end, this paper chose seven feature selection methods, namely, PCA, FA, MI, LASSO, KPCA, LR, and EN. RF was used as a classifier, and its results were tested using LOOCV. The optimal feature extraction method was selected by comparing the evaluation indicators. The results obtained by different feature selection methods are shown in Table 8.

Table 8. Evaluation results of different dimension reduction methods.

Method	ACC (%)	Sn (%)	Sp (%)	BACC (%)
PCA	84.44	88.64	80.24	84.44
FA	93.26	94.74	91.77	93.26
MI	93.29	96.93	89.65	89.14
LASSO	94.02	97.53	90.52	94.02
KPCA	86.48	88.36	84.59	86.48
LR	96.72	98.50	94.95	96.72
EN	97.28	98.78	95.78	97.28

Table 8 shows that different feature selection methods had a certain impact on the prediction performance of the model. On the training dataset, the BACC of the EN algorithm was 97.28%, which is 0.11%, 0.16%, 8.14%, 0.12%, 0.10%, and 0.52% higher than that of PCA, FA, MI, LASSO, KPCA, and LR. In addition, ACC, Sn, and SP were 97.28%, 98.78%, and 95.78%, respectively. This shows that, compared with PCA, FA, MI, LASSO, KPCA, and LR, the EN algorithm could effectively select a subset of features, and the prediction effect in the model was better.

From Figure 2, we can see the details of the ROC curve and PR curve of the training dataset under the seven feature selection methods. AUC and AUPR are important evaluation indicators that indicate the classification performance of the model. Higher values of AUC and AUPR denote the better the performance of the model. Results show that the area under the ROC and PR curves predicted by the EN algorithm was the largest. The AUC value was 0.9972, which is 0.02%–14.65% higher than the AUC value obtained by other feature selection methods. The AUPR value was 0.9974, which is 0.02%–6.10% higher than the AUPR value obtained by other feature selection methods. In conclusion, the feature selection result of the EN algorithm was the highest.

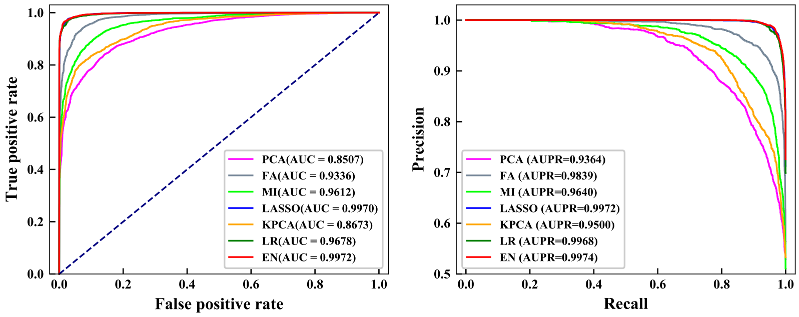

Figure 2. Comparison of receiver operating characteristic (ROC) and precision recall (PR) curves of seven dimensionality reduction methods. (**left**) ROC curve of the seven dimensionality reduction methods; (**right**) PR curve of the seven dimensionality reduction methods.

3.5. The Effect of The Classifier Algorithm on the Prediction Results

Classification algorithms play an important role in building the ECM protein prediction model. In this paper, six common classification algorithms were selected for comparison, namely GBDT, AdaBoost, naïve Bayes, SVM, LR, and RF. The feature subset constructed by EN was input into the six classifiers, and the results were tested by LOOCV. The prediction results of the training dataset on the six classifiers are shown in Table 9. The ROC and PR curves of the six classification algorithms are shown in Figure 3.

Table 9. Comparison of prediction results of six classifiers.

Classifier	ACC (%)	Sn (%)	Sp (%)	BACC (%)
GBDT	83.12	86.62	79.61	83.11
AdaBoost	93.67	95.30	92.05	93.68
Naïve Bayes	83.38	89.20	77.55	83.38
SVM	84.28	86.55	82.01	84.28
LR	91.18	93.97	88.39	91.18
RF	97.28	98.78	95.78	97.28

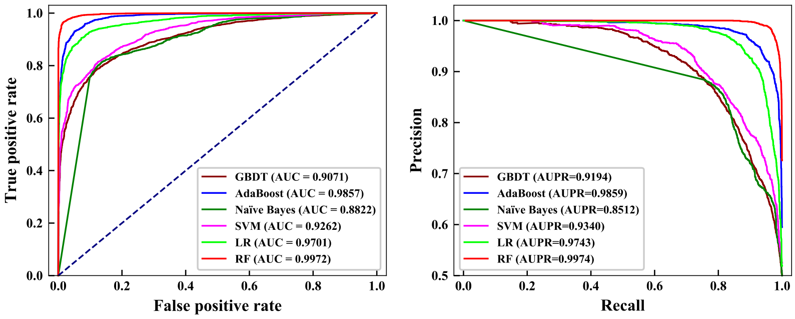

Figure 3. Comparison of ROC and PR curves of the six classifier algorithms. (**left**) ROC curve of the six classifier algorithms; (**right**) PR curve of the six classifier algorithms.

Table 9 shows that, on the training dataset, the BACC obtained by GBDT, AdaBoost, naïve Bayes, SVM, LR, and RF was 83.11%, 93.68%, 83.38%, 84.28%, 91.18%, and 97.28%. Among them, RF's BACC was 14.17%, 3.60%, 13.90%, 13.00%, and 6.10% higher than that of GBDT, AdaBoost, naïve Bayes, SVM, and LR algorithms. This shows that the classification performance of RF was superior.

It can be seen from Figure 3 that the AUC value of GBDT was 0.9071, the AUC value of AdaBoost was 0.9857, the AUC value of naïve Bayes was the lowest at 0.8822, the AUC value of SVM was 0.9262, the AUC value of LR was 0.9701, and the highest AUC value was achieved by RF at 0.9972. At the same time, the AUPR value of RF also reached 0.9974, which is 1.15%–14.62% higher than that of the other classifiers. Therefore, this paper chose RF as the classifier.

3.6. Comparison with Other Methods

At present, many researchers conducted prediction research on ECM proteins. In order to prove the validity and discrimination of the model, this section compares the ECMP-RF prediction model with other prediction models using the same datasets. To be fair, we use the LOOCV test to compare the expected method with other latest methods on the training dataset and testing dataset.

Figure 4 visually represents the prediction results of ECMP-RF and existing methods on the training dataset. Obviously, of all the methods, ECMPRED had the lowest ACC, Sp, and BACC, while PECMP [24] had the lowest Sn. It is worth noting that, although ECMPP [23] had the highest Sp, its Sn was poor, resulting in poor BACC results. At the same time, the proposed method ECMP-RF could improve the prediction performance of the model and get better results. As the most important evaluation index of the model, BACC was 97.3% in the training dataset. The BACC value of this paper was 19.5% higher than ECMPP, 24.2% higher than PECMP, 26.3% higher than ECMPERD [25], 10.9% higher than IECMP [27], 6.4% higher than ECMP-HybKNN [28], and 3.6% higher than TargetECMP [29]. In addition, the ACC and Sn of ECMP-RF had certain advantages. The comparison results of the training datasets are shown in Table S1 (Supplementary Materials). In short, we conclude that the method we propose is superior to existing models.

In order to objectively evaluate the methods proposed in this paper, we used the same testing dataset to test the prediction performance of the above methods. The specific results are shown in Figure 5. The BACC of the testing dataset on ECMP-RF was 97.9%, which is 33.9%, 29.2%, 42.9%, 20.4%, and 11.4% higher than ECMPP [23], PECMP [24], ECMPRED [25], IECMP [27], and

TargetECMP [29], respectively. The prediction results show that, compared with other ECM protein prediction methods, the ACC of the ECMP-RF model prediction was much higher than that of other methods, which verifies the excellent performance of ECMP-RF. The comparison results on the testing dataset are presented in Table S2 (Supplementary Materials).

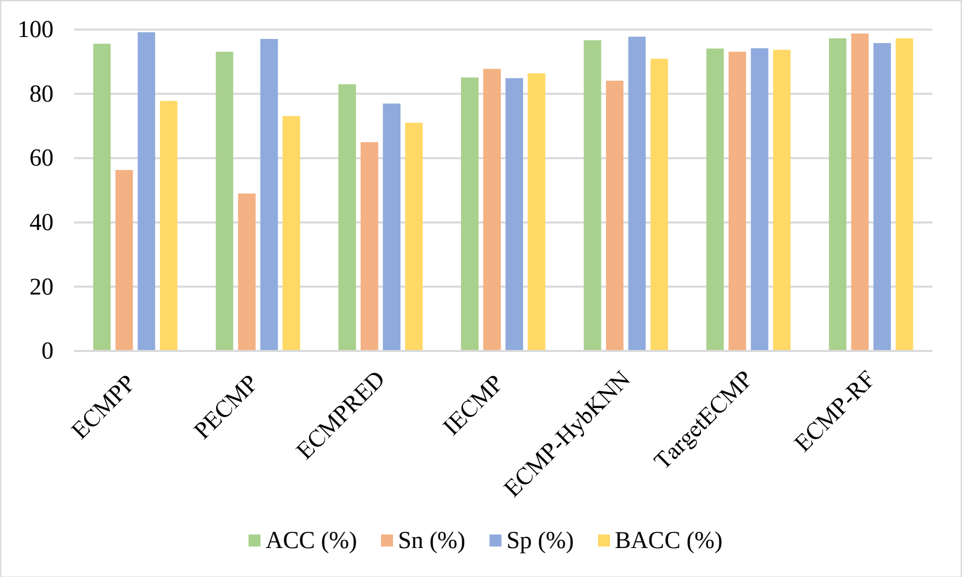

Figure 4. Comparison of ECMP-RF on training dataset with existing methods.

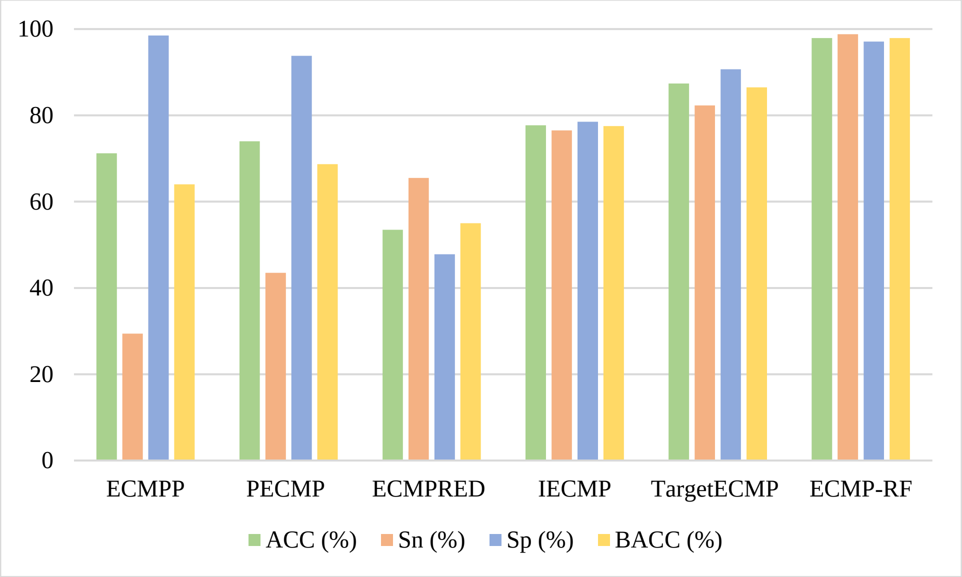

Figure 5. Comparison of ECMP-RF on testing dataset with existing methods.

4. Discussion

This paper built an ECMP-RF model to predict ECM proteins. PseAAC reflects the sequence information of ECM proteins. EBGW, LD, and AD can effectively extract the physicochemical properties of ECM proteins. PsePSSM extracts evolution information. Feature information is fused, and the SMOTE algorithm is used to solve the problem of data category imbalance. The balanced eigenvector is input into the EN for noise reduction. Using the best feature subset as the input of the RF to predict, ECMP-RF achieves 97.3% BACC in the training dataset through the most rigorous LOOCV, 3.6%–26.3% higher than the existing methods. It also achieves 97.9% BACC in the testing dataset through prediction, 11.4%–42.9% higher than other ECM prediction methods.

1. Multi-information fusion can extract protein sequence information more comprehensively and describe protein sequence features in detail compared with single feature extraction method.

2. The SMOTE algorithm is used for the first time to solve the problem of data category imbalance and improve the generalization ability of the model.

3. Compared with other dimensionality reduction methods, the EN algorithm can effectively remove the redundant information after fusion, and effectively retain the important feature information of ECM protein recognition.

4. The RF algorithm used in this paper is an ensemble classification algorithm, which can avoid overfitting and has good prediction performance in proteomics.

Although ECMP-RF improves the accuracy of ECM protein prediction compared with other methods, there is still room for improvement. In future work, we will use deep learning methods to further study ECM protein research, in order to accelerate the speed of model operation and improve the accuracy of the model prediction.

5. Conclusions

ECM proteins participate in a variety of biological processes, and they play an important role in cell life activities. A new prediction method ECMP-RF was proposed in this paper to recognize ECM proteins based on machine learning. The sequence information, evolution information, and physicochemical properties of ECM proteins were considered at the same time. SMOTE was used to solve the problem of class imbalance and avoid prediction bias. EN could effectively remove redundant features while keeping the important features of ECM proteins. The best feature subset was classified and predicted by RF in order to avoid overfitting. Compared with other methods, ECMP-RF advances the research of ECM proteins to a new phase.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1: Figure S1: Change of BACC value corresponding to different parameters of the feature extraction algorithm; Table S1: Comparison of ECMP-RF on training dataset with existing methods; Table S2: Comparison of ECMP-RF on testing dataset with existing methods.

Author Contributions: M.W., L.Y., and B.Y. conceptualized the algorithm, prepared the datasets, carried out experiments, and drafted the manuscript. X.C., C.C., and H.Z. carried out the data coding and data interpretation. Q.M. designed and analyzed the experiments and provided some suggestions for the manuscript writing. All authors read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Nature Science Foundation of China (Nos. 61863010 and 11771188), the Key Research and Development Program of Shandong Province of China (2019GGX101001), and the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007). This work used the Extreme Science and Engineering Discovery Environment, which is supported by the National Science Foundation (No. ACI-1548562).

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

References

1. Campbell, N.E.; Kellenberger, L.; Greenaway, J.; Moorehead, R.A.; Linnerth-Petrik, N.M.; Petrik, J. Extracellular mtrix proteins and tumor angiogenesis. *J. Oncol.* **2010**, *2010*, 586905.
2. Barkan, D.; Green, J.E.; Chambers, A.F. Extracellular matrix: A gatekeeper in the transition from dormancy to metastatic growth. *Eur. J. Cancer* **2010**, *46*, 1181–1188.
3. Liotta, L.A. Tumor invasion and extracellular matrix. *Lab. Investig.* **1983**, *49*, 636–649.
4. Adams, J.C.; Watt, F.M. Regulation of development and differentiation by the extracellular matrix. *Development* **1993,** *117*, 1183–1198.
5. Mathews, S.; Bhonde, R.; Gupta, P.K.; Totey, S. Extracellular matrix protein mediated regulation of the osteoblast differentiation of bone marrow derived human mesenchymal stem cells. *Differentiation* **2012**, *84*, 185–192.
6. Endo, Y.; Ishiwata-Endo, H.; Yamada, K. Extracellular matrix protein anosmin promotes neural grest formation and regulates FGF, BMP, and WNT activities. *Dev. Cell* **2012**, *23*, 305–316.

7. Kim, S.-H.; Turnbull, J.; Guimond, S. Extracellular matrix and cell signalling: The dynamic cooperation of integrin, proteoglycan and growth factor receptor. *J. Endocrinol*. **2011**, *209*, 139–151.
8. Aitken, K.J.; Bägli, D.J. The bladder extracellular matrix. Part I: Architecture, development and disease. *Nat. Rev. Urol.* **2009**, *6*, 596–611.
9. Karsdal, M.A.; Nielsen, M.J.; Sand, J.M.; Henriksen, K.; Genovese, F.; Bay-Jensen, A.C.; Victoria, S.; Adamkewicz, J.I.; Christiansen, C.; Leeming, D.J. Extracellular matrix remodeling: The common denominator in connective tissue diseases possibilities for evaluation and current understanding of the matrix as more than a passive architecture, but a key player in tissue failure. *Proteins* **2012**, *80*, 1522–2544.
10. Cromar, G.L.; Xiong, X.; Chautard, E.; Ricard-Blum, S.; Parkinson, J. Toward a systems level view of the ECM and related proteins: A framework for the systematic definition and analysis of biological systems. *Proteins* **2012**, *80*, 1522–1544.
11. Fallon, J.R.; McNally, E.M. Non-Glycanated Biglycan and LTBP4: Leveraging the extracellular matrix for Duchenne Muscular Dystrophy therapeutics. *Matrix Biol.* **2018**, 68-69, 616-627.
12. Ma, F.; Tremmel, D.M.; Li, Z.; Lietz, C.B.; Sackett, S.D.; Odorico, J.S.; Li, L. In depth quantification of extracellular matrix proteins from human pancreas. *J. Proteome Res*. **2019**, *18*, 3156–3165.
13. Gonzalez-Pujana, A.; Santos-Vizcaino, E.; García-Hernando, M.; Hernaez-Estrada, B.; de Pancorbo, M.M.; Benito-Lopez, F.; Igartua, M.; Basabe-Desmonts, L.; Hernandez, R.M. Extracellular matrix protein microarray-based biosensor with single cell resolution: Integrin profiling and characterization of cell-biomaterial interactions. *Sens. Actuators B Chem*. **2019**, *299*, 126954.
14. Li, Z.C.; Lai, Y.H.; Chen, L.L.; Chen, C.; Xie, Y.; Dai, Z.; Zou, X.Y. Identifying subcellular localizations of mammalian protein complexes based on graph theory with a random forest algorithm. *Mol. BioSyst*. **2013**, *9*, 658–667.
15. Chen, X.; Chen, M.; Ning, K. BNArray: An R package for constructing gene regulatory networks from microarray data by using Bayesian network. *Bioinformatics* **2006**, *22*, 2952–2954.
16. Tang, Y.R.; Chen, Y.Z.; Canchaya, C.A.; Zhang, Z. GANNPhos: A new phosphorylation site predictor based on a genetic algorithm integrated neural network. *Protein Eng. Des. Sel*. **2007**, *20*, 405–412.
17. Yamada, K.D.; Omori, S.; Nishi, H.; Miyagi, M. Identification of the sequence determinants of protein N-terminal acetylation through a decision tree approach. *BMC Bioinform.* **2017**, *18*, 289.
18. Ahmad, K.; Waris, M.; Hayat, M. Prediction of Protein Submitochondrial Locations by Incorporating Dipeptide Composition into Chou's General Pseudo Amino Acid Composition. *J. Membr. Biol*. **2016**, *249*, 293–304.
19. Chen, T.Q.; 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.
20. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of online learning and an application to boosting. *J. Comput. Syst. Sci*. **1997**, *55*, 119–139.
21. Wang, B.; Wang, X.F.; Howell, P.; Qian, X.; Huang, K.; Riker, A.I.; Ju, J.; Xi, Y. A personalized microRNA microarray normalization method using a logistic regression model. *Bioinformatics* **2010**, *26*, 228–234.
22. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat*. **2001**, *29*, 1189–1232.
23. Jung, J.; Ryu, T.; Hwang, Y.; Lee, E.; Lee, D. Prediction of extracellular matrix proteins based on distinctive sequence and domain characteristics. *J. Comput. Biol*. **2010**, *17*, 97–105.
24. Anitha, J.; Rejimoan, R.; Sivakumar, K.C.; Sathish, M. Prediction of extracellular matrix proteins using SVM^{hmm} classifier. *IJCA Spec. Issue Adv. Comput. Commun. Technol. HPC Appl.* **2012**, *1*, 7–11.
25. Kandaswamy, K.K.; Pugalenthi, G.; Kalies, K.U.; Hartmann, E.; Martinetz, T. EcmPred: Prediction of extracellular matrix proteins based on random forest with maximum relevance minimum redundancy feature selection. *J. Theor. Biol.* **2013**, *317*, 377–383.
26. Zhang, J.; Sun, P.; Zhao, X.; Ma, Z. PECM: Prediction of extracellular matrix proteins using the concept of chou's pseudo amino acid composition. *J. Theor. Biol.* **2014**, *363*, 412–418.
27. Yang, R.; Zhang, C.; Gao, R.; Zhang, L. An ensemble method with hybrid features to identify extracellular matrix proteins. *PLoS ONE* **2015**, *10*, e0117804.
28. Ali, F.; Hayat, M. Machine learning approaches for discrimination of extracellular matrix proteins using hybrid feature space. *J. Theor. Biol.* **2016**, *403*, 30–37.

29. Kabir, M.; Ahmad, S.; Iqbal, M.; Swati, Z.N.; Liu, Z.; Yu, D.J. Improving prediction of extracellular matrix proteins using evolutionary information via a grey system model and asymmetric under-sampling technique. *Chemom. Intell. Lab*. **2018**, *174*, 22–32.
30. David, C.C.; Jacobs, D.J. Principal component analysis: A method for determining the essential dynamics of proteins. *Methods Mol. Biol.* **2013**, *1084*, 193–226.
31. Engemann, D.A.; Gramfort, A. Automated model selection in covariance estimation and spatial whitening of MEG and EEG signals. *NeuroImage* **2015**, *108*, 328–342.
32. Tabbaa, O.P.; Jayaprakash, C. Mutual information and the fidelity of response of gene regulatory models. *Phys. Biol.* **2014**, *11*, 046004.
33. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429.
34. Li, J.; Li, X.; Tao, D. KPCA for semantic object extraction in images. *Pattern Recogn*. **2008**, *41*, 3244–3250.
35. Hsieh, F.Y.; Bloch, D.A.; Larsen, M.D. A simple method of sample size calculation for linear and logistic regression. *Stat. Med.* **1998**, *17*, 1623–1634.
36. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *2*, 301–320.
37. You, Z.H.; Zhu, L.; Zheng, C.H.; Yu, H.J.; Deng, S.P.; Ji, Z. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinform.* **2014**, *15*, S9.
38. Zhang, Z.H.; Wang, Z.H.; Zhang, Z.R.; Wang, Y.X. A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. *FEBS Lett.* **2006**, *580*, 6169–6174.
39. Wang, X.; Yu,B.; Ma, A.; Chen, C.; Liu, B.; Ma, Q. Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* **2019**, *35*, 2395–2402.
40. Tian, B.G.; Wu, X.; Chen, C.; Qiu, W.Y.; Ma, Q.; Yu, B. Predicting protein–protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach. *J. Theor. Biol.* **2019**, *462*, 329–346.
41. Zhou, H.Y.; Chen, C.; Wang, M.H.; Ma, Q.; Yu, B. Predicting Golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion. *IEEE Access* **2019**, *7*, 144154–144164.
42. Yu, B.; Qiu, W.; Chen, C.; Ma, A.; Jiang, J.; Zhou, H.; Ma, Q. SubMito-XGBoost: Predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics* **2019**, doi: 10.1093/bioinformatics/btz734.
43. Chou, K.C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **2001**, *43*, 246–255.
44. Cui, X.W.; Yu, Z.M.; Yu, B.; Wang, M.H.; Tian, B.G.; Ma, Q. UbiSitePred: A novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components. *Chemom. Intell. Lab.* **2019**, *184*, 28–43.
45. Yu, B.; Lou, L.; Li, S.; Zhang, Y.; Qiu, W.; Wu, X.; Wang, M.; Tian, B. Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising. *J. Mol. Graph. Model.* **2017**, *76*, 260–273.
46. Butt, A.H.; Rasool, N.; Khan, Y.D. Prediction of antioxidant proteins by incorporating statistical moments based features into Chou's PseAAC. *J. Theor. Biol.* **2019**, *473*, 1–8.
47. Yu, B.; Li, S.; Qiu, W.Y.; Wang, M.H.; Du, J.W.; Zhang, Y.S.; Chen, X. Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction. *BMC Genom.* **2018**, *19*, 478.
48. Yu, B.; Li, S.; Qiu, W.Y.; Chen, C.; Chen, R.X.; Wang, L.; Wang, M.H.; Zhang, Y. Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising. *Oncotarget* **2017**, *8*, 107640–107665.
49. Yu, B.; Li, S.; Chen, C.; Xu, J.; Qiu, W.Y.; Chen, R. Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition. *Chemom. Intell. Lab.* **2017**, *167*, 102–112.
50. Cheng, X.; Xiao, X.; Chou, K.C. pLoc_bal-mPlant: Predict subcellular localization of plant proteins by general PseAAC and balancing training dataset. *Curr. Pharm. Des.* **2018**, *24*, 4013–4022.
51. Lin, H.; Wang, H.; Ding, H.; Chen, Y.L.; Li, Q.Z. Prediction of subcellular localization of apoptosis protein using chou's pseudo amino acid composition. *Acta Biotheor.* **2009**, *57*, 321–330.

52. Jiao, Y.S.; Du, P.F. Predicting Golgi-resident protein types using pseudo amino acid compositions: Approaches with positional specific physicochemical properties. *J. Theor. Biol.* **2016**, *391*, 35–42.
53. Shen, H.B.; Chou, K.C. Nuc-PLoc: A new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Eng. Des. Sel.* **2007**, *20*, 561–567.
54. Qiu, W.Y.; Li, S.; Cui, X.W.; Yu, Z.M.; Wang, M.H.; Du, J.W.; Peng, Y.J.; Yu, B. Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition. *J. Theor. Biol.* **2018**, *450*, 186–103.
55. Shi, H.; Liu, S.; Chen, J.; Li, X.; Ma, Q.; Yu, B. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics* **2019**, *111*, 1839–1852.
56. Chen, C.; Zhang, Q.M.; Ma, Q.; Yu, B. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemom. Intell. Lab.* **2019**, *191*, 54–64.
57. Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T.T.; Wang, Y.; Webb, G.I.; Smith, A.I.; Daly, R.J.; Chou, K.C.; et al. iFeature: A python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* **2018**, *34*, 2499–2502.
58. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2012**, *16*, 321–357.
59. Blagus, R.; Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* **2013**, *14*, 106.