Computers, Materials & Continua
DOI:10.32604/cmc.2021.017769
Article

Tech Science Press

DTLM-DBP: Deep Transfer Learning Models for DNA Binding Proteins Identification

Sara Saber[1], Uswah Khairuddin[2,*], Rubiyah Yusof[2] and Ahmed Madani[1]

[1]Department of Computer Engineering, Faculty of Engineering, Arab Academy of Science and Technology, Egypt
[2]Centre for Artificial Intelligence & Robotics, Malaysia-Japan International Institute of Technology,
Universiti Teknologi Malaysia
[*]Corresponding Author: Uswah Khairuddin. Email: uswah.kl@utm.my
Received: 10 January 2021; Accepted: 16 March 2021

Abstract: The identification of DNA binding proteins (DNABPs) is considered a major challenge in genome annotation because they are linked to several important applied and research applications of cellular functions e.g., in the study of the biological, biophysical, and biochemical effects of antibiotics, drugs, and steroids on DNA. This paper presents an efficient approach for DNABPs identification based on deep transfer learning, named "DTLM-DBP." Two transfer learning methods are used in the identification process. The first is based on the pre-trained deep learning model as a feature's extractor and classifier. Two different pre-trained Convolutional Neural Networks (CNN), AlexNet 8 and VGG 16, are tested and compared. The second method uses the deep learning model as a feature's extractor only and two different classifiers for the identification process. Two classifiers, Support Vector Machine (SVM) and Random Forest (RF), are tested and compared. The proposed approach is tested using different DNA proteins datasets. The performance of the identification process is evaluated in terms of identification accuracy, sensitivity, specificity and MCC, with four available DNA proteins datasets: PDB1075, PDB186, PDNA-543, and PDNA-316. The results show that the RF classifier, with VGG-Net pre-trained deep transfer learning features, gives the highest performance. DTLM-DBP was compared with other published methods and it provides a considerable improvement in the performance of DNABPs identification.

Keywords: DNABPs; deep transfer learning; AlexNet 8; VGG 16; SVM; RF

1 Introduction

Deoxyribonucleic Acid (DNA) represents the cell blueprint that contains the main information that codes all organisms. DNA can perform its functions with the help of thousands of proteins, which are called DNA binding proteins (DNABPs). DNABPs have several jobs, such controlling protein production, regulating cell growth and storing DNA in the nucleus. DNABPs play an important role in the structural composition of DNA. In addition, they regulate and control different cellular processes such as the transcription, replication, recombination, repair and modification of DNA.

DNABPs identification is considered a major challenge of genome annotation because they have several linked cellular functions. The identification process may include: identifying the DNABPs (positive sample) from the non-DNABPs (negative sample) [1], identifying the single-stranded DNABPs from the double-stranded DNABPs [2], or identifying the DNABPs from the Ribonucleic acid-binding proteins (RNABPs) [3–5]. In this paper, the identification process is formulated as a binary classification problem to identify DNABPs and non-DNABPs. DNABPs are the proteins that have DNA binding domains and they generally interact with the major groove of B-DNA. Non-DNABPs, on the other hand, are the structural proteins within the chromosomes.

Several experimental technical methods can be used for identifying DNABPs, but they are time-consuming and expensive [6]. Therefore, there is a significant need to find a suitable and efficient computational method for replacing these experimental methods. Recently, several computational and statistical methods have been proposed for DNABPs identification, but most of these methods cannot provide the invaluable knowledge base for DNABPs identification. With the advancements in machine and deep learning techniques over recent years, several methods based on machine and deep learning have been presented.

Zhu et al., proposed a method for DNABPs identification based on the position-specific scoring matrices (PSSM) and co-occurrence matrix. The results achieved an accuracy of 97.06% for Yeast dataset, 98.95% for Human dataset, and 89.69% for H.Pylori dataset [7]. The PSSM with SVM (PSFM-DT) tested by Xu et al. [8] achieved an accuracy of 79.96% for PDB1075 dataset, and 79.96% for PDB186 dataset. In addition, the PSSM with RF tested by Waris et al. [9] achieved an accuracy of 92.3% for their tested dataset. Chowdhury et al., proposed a method (iDNAProt-ES) for DNABPs identification by extracting the structural and evolutionary features that feed the SVM predictor. The results achieved an accuracy of 90.18% for the jack-knife dataset [10]. Xu used the random forest for DNABPs identification. The results achieved an accuracy of 85.57 for the jack-knife dataset [11].

Zhang et al., proposed a method for DNABPs identification by combining the position-specific frequency matrix and the distance-bigram transformation (PSFM-DBT). The results achieved an accuracy of 81.02% for PDB1075 dataset, and 80.65% for PDB186 dataset [12]. Zhang et al., made features with a fusion of evolutionary, structural, and physicochemical features for DNABPs identification, and used the binary firefly optimization for removing the redundant features. The results achieved an accuracy of 91% for the DNA dataset, and 0.80.9% for PDB186 dataset [13]. Ma et al., proposed a method for DNABPs identification based on selecting the hybrid features using the random forest. The results achieved an accuracy of 89.56% for Mainsett dataset [14].

Moreover, Shen et al., used the multi-scale local average blocks approach for DNABPs identification. The results achieved an accuracy of 91.80% for PDNA-543 dataset, 92.06% for PDNA-41 dataset, 90.23% for PDNA-316 dataset, and 77.6% for PDNA-52 dataset [15]. Krishna et al., proposed a DNABPs identification (DNA-Prot) method by incorporating the evolutionary features into the pseudo-amino acid composition. The results achieved an accuracy of 81.83% for DNA-Prot dataset, and 61.42% for DNA binder dataset [16]. This method was modified by adding the grey model and named iDNA-Prot [17]. Fu et al., applied the same method on the jack-knife test and independent test, which achieved an accuracy of 89.77% and 88.71%, respectively [18].

Wei et al. [19] used RF in the training, called the Method by Local-DPP model. Moreover, Liu et al. [20] used SVM and called it iDNAPro-PseAAC. This method was improved through dimension reduction by Liu et al. [21] and renamed iDNA-Pro-dis. The concept of Pse-AAC was applied in other models called DNABinder [22], PseDNA-Pro [23], and DPP-PseAAC [24]. Biological information was added by Zaman et al. [25] and named (HMMBinder).

Szilagyi et al. presented a method for DNABPs identification (DNABIND) based on the amino acid proportions in the sequence of the protein. The results achieved an accuracy of 67.70% for PDB186 dataset [26]. Gao et al. presented a threading-based method for DNABPs identification (DNA-Threader). The results achieved an accuracy of 59.7% for PDB186 dataset [27]. Szilagyi et al. presented a DNABPs identification method (DNABIND) based on hybrid feature selection using RF and Gaussian naive Bayes (DBPPred). The results achieved an accuracy of 76.90% for PDB186 dataset [28].

Zhang et al., proposed a DNABPs identification method using bootstrap multiple CNN. The results achieved an accuracy of 90.77% for PDNA-543 dataset and 91.04% for PDNA-316 dataset [29]. They used the long short-term memory and CNN. The results achieved an accuracy of 81.83% for DNA-Prot dataset, and 89.19% for Chip-seq dataset [30]. Liu et al., proposed a method for DNABPs identification by combining the auto-cross covariance with ensemble learning (iDNA-KACC). The results achieved an accuracy of 75.16% for the tested dataset [31]. Qu et al., proposed a method for DNABPs identification using mixed feature representation methods. The results achieved an accuracy of 77.43% for PDB1075 dataset, and 81.58% for PDB186 dataset [32]. Hu et al. [33] combined the sequence features with multiple SVMs and named the method TargetDNA. Si et al. [34] presented a meta-based DNABPs identification and named it MetaDBSite.

The main contribution of this paper is the testing and adaptation of pre-trained deep transfer learning models for DNABPs sequence identification. The paper presents a novel approach for DNABPs identification using deep transfer learning. In this approach, two transfer learning methods were tested and compared; in the first method, the pre-trained deep CNN (AlexNet 8 or VGG 16) learning model was used as a feature's extractor and classifier. In the second method, the deep learning model was used as a feature's extractor only, while the classifier was either the SVM or RF. The proposed approach was tested using different DNA proteins datasets. The performance of the identification process was evaluated in terms of identification accuracy, sensitivity, specificity, and MCC with four available DNA proteins datasets: PDB1075, PDB186, PDNA-543, and PDNA-316 datasets. The results show that the RF classifier with VGG-Net pre-trained deep transfer learning features produced the highest performance. DTLM-DBP was compared with the other published methods and found to represent a considerable improvement in the performance of DNABPs identification. The remainder of the paper is organized as follows: the second section will present the proposed methodology, the third section gives the results, and the conclusion will be given in the last section.

2 Materials and Methods

The general block diagram of the DNABPs identification process in this paper is shown in Fig. 1. Two transfer learning methods were carried out. In the first method, the protein sequences were adapted to CNN models using 1D convolutions layers, then one of the pre-trained deep CNN learning models was used as a feature's extractor and classifier. In the second method, the deep learning model was used as a feature's extractor only, while the classifier was either the SVM or RF. More details about each block will be presented in the following subsections.

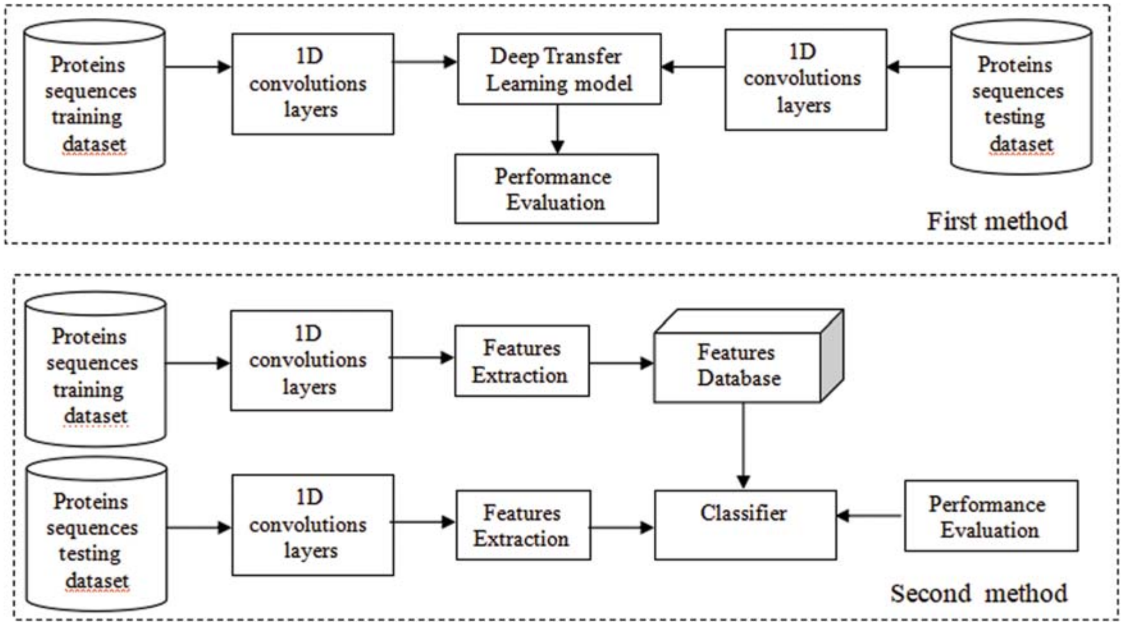

Figure 1: DNABPs identification process

2.1 Datasets

There are several publicly available protein sequences datasets, most of which were collected from the protein data bank (PDB). The researchers collected the sequences data from the PDB website by searching for words such as 'DNA binding,' 'DNA protein' and other related terms, Then, certain processing procedures were undertaken to avoid the inclusion of redundant data, and finally, the obtained datasets were used in the research. To guarantee the reliability of the proposed approach and for performance evaluation comparison purposes, pre-collected publicly available datasets were used that had been used by several researchers in the literature.

The experimental work was implemented on four different DNABPs datasets: PDB1075, PDB186, PDNA-543, and PDNA-316 datasets. PDB1075 dataset was collected by Liu et al. [21], and included 1,075 protein samples; 525 samples were positive DNABPs and 550 samples were negative non-DNABPs. PDB186 dataset was collected by Lou et al. [28], and included 186 protein samples; 93 samples were positive DNABPs and 93 samples were negative non-DNABPs. PDNA-543 dataset was collected by Hu et al. [33], and included 144,544 protein samples; 9,549 samples were positive DNABPs and 134,995 samples were negative non-DNABPs. PDNA-316 dataset was collected by Si et al. [34], and included 72,718 protein samples; 5,609 samples were positive DNABPs and 67,109 samples were negative non-DNABPs.

2.2 Deep Transfer Learning Models

In this paper, two pre-trained deep transfer learning models, AlexNet and VGG-Net, were adapted for the identification of DNABPs sequences. The model architecture of each training model will be presented. These two models had been selected from the large number of pre-trained deep learning transfer models because, according to the literature, they are the most successful models in terms of the identification process, while their architectures are simple and contain different numbers of convolution layers.

2.2.1 AlexNet-8 Pre-Trained Deep Transfer Learning Model

AlexNet-8 is a CNN that is 8 layers deep, and was introduced by Krizhevsky et al. [35]. The number of parameters in AlexNet-8 is 60 million and the number of neurons is 650,000. It consists of 8 layers (5 convolutional and 3 fully connected), as shown in the model architecture in Fig. 2 [36]. The first and second convolutional layers are followed by normalization and a max-pooling layer, the third and fourth convolutional layers are connected directly, and the last convolutional layer is followed by a max-pooling layer. The output of the convolutional layer passes through a series of two fully connected layers, in which the second fully connected layer is fed into the SoftMax classifier.

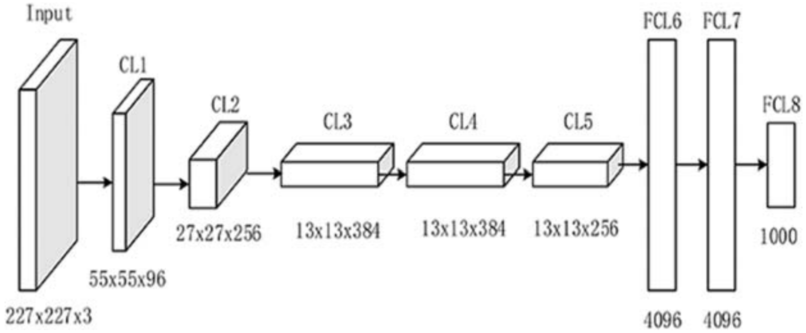

Figure 2: AlexNet-8 model architecture

2.2.2 VGG-16 Pre-Trained Deep Transfer Learning Model

VGG-16 is a CNN model which is 16 layers deep, and was introduced by Simonyan and Zisserman in 2014 [37,38]. According to the literature, VGG-16 offers a considerable improvement over AlexNet in several applications because it is rich with several feature representations that can be used for a wide range of applications. The VGG-16 model architecture is shown in Fig. 3 [39]. It consists of a 16-layer network comprised of convolutional layers.

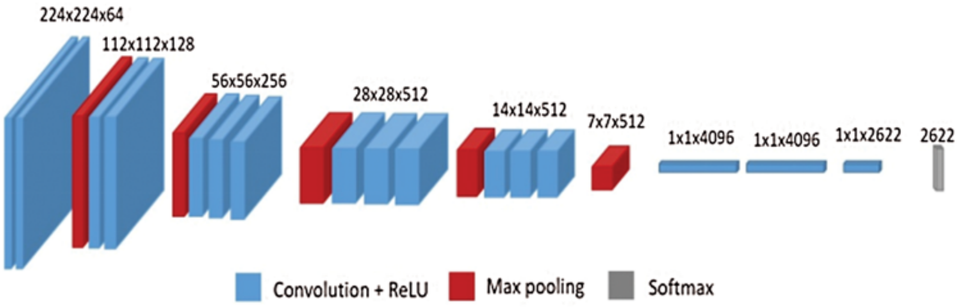

Figure 3: VGG-16 model architecture

2.3 Classifiers

The DNA proteins identification process is mainly a binary classification problem between two classes. The first class is the DNABPs that have DNA binding domains and interact with the DNA. The second class is the non-DNABPs, such as the structural proteins within the chromosomes. Several classifiers are suitable for binary classification; the most commonly used classifiers for DNA proteins identification are SVM [8,22,33] and RF [14,16,17,28]. In this paper, the two classifiers were used and compared.

2.3.1 SVM Classifier

SVM is a set of related supervised-learning models introduced by Cortes et al. [40]. It minimizes the identification error and maximizes the geometric margin. SVMs are the most suitable binary linear identification methods [40–43]. SVM works for two-class problems by separating the data by a separating hyperplane, as shown in Fig. 4.

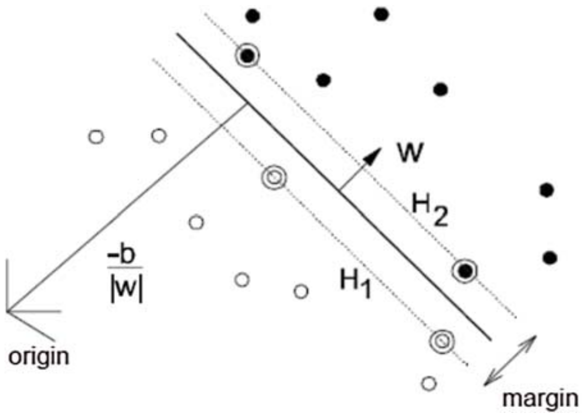

Figure 4: SVM separating hyperplanes

In Fig. 4, consider that the training sequences are represented by {x_i, y_i}, i = 1, . . . , l, $y_i = \pm 1$, $x_i \in R_d$, x points lie on the hyperplane and satisfy the condition x. w + b = 0, w a is normal to the hyperplane. This can be formulated as [44]:

$$x_i.w + b \geq +1 \quad \text{for } y_i = +1 \tag{1}$$

$$x_i.w + b \leq -1 \quad \text{for } y_i = -1 \tag{2}$$

The primal Lagrange is given as [44]:

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i (x_i.w + b) - 1) \tag{3}$$

where αi, $i = 1, \ldots, l$ are the positive Lagrange multipliers, $\|w\|$ is the Euclidean norm of w.

For minimizing L_P with respect to *w, b*, using the conditions:

$$\frac{\delta L_P}{\delta w_0} = 0 \quad \text{gives } w_0 = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{4}$$

$$\frac{\delta L_P}{\delta b_0} = 0 \quad \text{gives } b_0 = \sum_{i=1}^{l} \alpha_i y_i \tag{5}$$

Using Eqs. (3)–(5), the dual Lagrangian will be:

$$L_d(\alpha) = \sum \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j x_i x_j \tag{6}$$

The mapping of training vectors xi into the higher dimensional space uses a function called kernel function $K(x_i, x_j) \equiv \Phi(x_i)\Phi(x_j)$. There are several SVMs kernel functions, such as:

Linear kernel:

$$K(x_i, x_j) = x_i.x_j \tag{7}$$

Polynomial kernel:

$$K(x_i, x_j) = (\gamma x_i.x_j + r)d, \quad \gamma > 0 \tag{8}$$

RBF kernel:

$$K(x_i, x_j) = exp(-\gamma \left\| x_i - x_j \right\| 2), \quad \gamma > 0 \tag{9}$$

Sigmoid kernel:

$$K(x_i, x_j) = tanh(\gamma x_i.x_j + r) \tag{10}$$

where γ, r and d are kernel parameters.

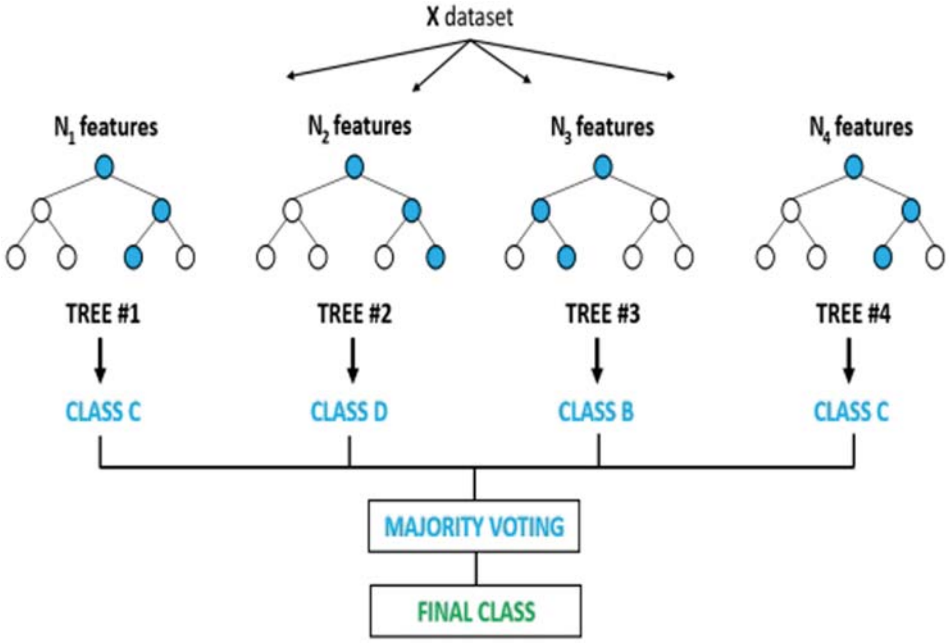

Figure 5: RF algorithm

The DNA protein sequences identification was carried out using the SVM Matlab Toolbox with different kernel functions: linear, polynomial, RBF, and sigmoid kernel. DNABPs identification using SVM can be carried out in two steps. The first step is building the identification simulation model, while the second step is the feature matching for the model performance evaluation. In the modelling step, the features related to the DNA protein sequences are stored. When a tested sequence arrives, its features are matched with the stored features in the model and the identification decision is taken based on the matching process.

2.3.2 RF Classifier

RF is a tree collection introduced by Ho [45]; each tree is grown through a subset of all the possible attributes of the input features vectors [46]. It constructs the decision ensemble in random trees based on the input features, and the final identification decision is obtained by combining the results from the trees via voting, as shown in Fig. 5.

3 Results and Discussions

3.1 Performance Evaluation Metrics

The performance of the DNA protein sequences identification system is normally evaluated using wide performance metrics, such as identification accuracy, sensitivity, specificity, and Matthew's correlation coefficient. These metrics can be calculated using four parameters obtained from the testing of the identification system with a certain dataset. The system tests the DNA protein sequences if it is a DNABP (positive sample) or non-DNABP (negative sample). For each DNABP testing, if the test result is positive, this means that the system identifies it as correct (True), and accumulating the positive true results for all the tested protein sequences in the dataset gives the T_p number. If the test result is negative, this means that the system identifies it as incorrect (False) and the accumulation gives the F_n number. For each non-DNABP testing, if the test result is positive, this means that the system identifies it as incorrect (False), and the accumulation gives the F_p number. If the test result is negative, this means that the system identifies it as correct (True) and the accumulation gives the T_n number. Using these four numbers, it is possible to calculate:

1. Accuracy

$$\text{Acc} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \times 100\% \tag{11}$$

2. Sensitivity

$$\text{Sen} = \frac{T_P}{T_P + F_N} \times 100\% \tag{12}$$

3. Specificity

$$\text{Acc} = \frac{T_N}{T_N + F_P} \times 100\% \tag{13}$$

4. Matthew's correlation coefficient

$$\text{MCC} = \frac{(T_P T_N) - (F_P F_N)}{\sqrt{(T_P + F_P)(T_P + F_N)(T_N + F_P)(T_N + F_N)}} \tag{14}$$

The accuracy, the sensitivity and the specificity are percentages, while the MCC ranges from -1 to $+1$; the perfect classifier should give 100% for the three first parameters and $+1$ MCC.

3.2 Deep Transfer Learning Models

This section presents the results of the first DNABPs identification method, which is based on the pre-trained deep transfer learning models as the features extractor and classifier. Two pre-trained deep transfer learning models, AlexNet and VGG-Net, were tested and compared in terms of identification accuracy, sensitivity, specificity, and MCC for the four examined DNA proteins datasets, as shown in Tab. 1.

Table 1: Performance comparison between deep transfer learning models

Dataset	Method	Acc (%)	Sen (%)	Spe (%)	MCC
PDB1075	AlexNet	93.02	91.51	94.56	0.86
	VGGNet	92.56	91.74	93.36	0.85
PDB186	AlexNet	75.27	82.19	70.79	0.52
	VGGNet	77.42	85.91	72.17	0.56
PDNA-543	AlexNet	86.01	30.21	98.80	0.45
	VGGNet	91.67	43.56	99.12	0.58
PDNA-316	AlexNet	88.23	36.31	97.26	0.45
	VGGNet	90.93	44.03	96.95	0.48

The results in Tab. 1 show that the VGG-Net 16 pre-trained deep transfer learning model gives higher performance than AlexNet. This may be because the 16-layer VGGnet is deeper than the 8-layer AlexNet, and the VGGnet is rich with several feature representations.

3.3 Classifiers Tuning

This section presents the results of the second DNABPs identification method, which is based on the pre-trained deep transfer learning models as the features extractor. The classifier is one of the two different classifiers (SVM or RF) used for the identification process. The identification accuracy, sensitivity, specificity, and MCC for the four examined DNA proteins datasets are as shown in Tab. 2.

The results in Tab. 2 show that the RF classifier with VGG-Net pre-trained deep transfer learning features gives the highest performance compared to the other approaches.

3.4 Performance Comparison with Existing Methods

The performance of the proposed DNABPs identification method (DTLM-DBP) was compared with the other published methods for the four available DNA proteins datasets: PDB1075, PDB186, PDNA-543, and PDNA-316 datasets. For PDB1075 dataset, DTLM-DBP was compared with DNAbinder [22], DNA-Prot [16], iDNA-Prot [17], iDNA-Prot-dis [21], PSSM-DT [8], PseDNA-Pro [23], iDNAPro-PseAAC [20], PSFM-DBT [12], Mixed Feature [32], Local-DPP [19], iDNAProt-ES [10], HMMBinder [25], iDNA-KACC [31], and DPP-PseAAC [24], as shown in Tab. 3.

The results in Tab. 3. show that the proposed method gives a better performance than the other published methods.

For PDB186 dataset, DTLM-DBP was compared with DNABIND [26], DNAbinder [22], DNA-Threader [27], DNA-Prot [16], DBPPred [28], iDNA-Prot [17], iDNA-Prot-dis [21], PSSM-DT [8], iDNAPro-PseAAC [20], Mixed Feature [32], PseDNA-Pro [23], iDNAProt-ES [10], PSFM-DBT [12], Local-DPP [19], HMMBinder [25], DPP-PseAAC [24], and iDNA-KACC-EL [31], as shown in Tab. 4. The results show the superiority of the proposed method over the other published methods.

Table 2: Performance comparison between SVM and RF classifiers

Dataset	Features	Classifier	Acc (%)	Sen (%)	Spe (%)	MCC
PDB1075	AlexNet	SVM	92.00	90.28	93.77	0.84
		RF	94.23	92.63	95.86	0.88
	VGGNet	SVM	92.28	90.62	93.97	0.85
		RF	96.34	94.83	97.94	0.93
PDB186	AlexNet	SVM	75.27	83.10	70.48	0.52
		RF	79.57	87.67	74.33	0.61
	VGGNet	SVM	76.88	84.72	71.93	0.55
		RF	81.18	91.43	75.00	0.64
PDNA-543	AlexNet	SVM	84.92	20.27	95.66	0.22
		RF	90.07	34.90	96.89	0.40
	VGGNet	SVM	87.07	24.29	95.87	0.27
		RF	93.05	48.11	97.57	0.53
PDNA-316	AlexNet	SVM	87.15	32.61	96.59	0.39
		RF	90.03	42.06	97.98	0.52
	VGGNet	SVM	88.69	35.63	96.28	0.39
		RF	93.38	55.35	97.69	0.60

Table 3: Comparison of DTLM-DBP with previous methods for PDB 1075 dataset

Method	Acc (%)	Sen (%)	Spe (%)	MCC
DNAbinder [22]	79.09	48.00	81.40	0.48
DNA-Prot [16]	72.55	82.67	59.76	0.44
iDNA-Prot [17]	75.40	83.81	64.73	0.50
iDNA-Prot-dis [21]	77.30	79.4	75.27	0.54
PSSM-DT [8]	79.96	81.91	78.00	0.62
PseDNA-Pro [23]	76.55	79.61	73.63	0.53
iDNAPro-PseAAC [20]	76.76	75.62	77.45	0.53
PSFM-DBT [12]	81.02	84.19	78.0	0.62
Mixed feature [32]	77.43	77.84	77.05	0.55
Local-DPP [19]	79.20	84.00	74.50	0.59
iDNAProt-ES [10]	90.18	90.38	90.00	0.80
HMMBinder [25]	86.33	87.00	85.50	0.72
iDNA-KACC [31]	75.16	77.52	72.90	0.50
DPP-PseAAC [24]	95.91	94.10	97.64	0.92
DTLM-DBP (proposed)	96.34	94.83	97.94	0.93

For PDNA-543 dataset, DTLM-DBP was compared with TargetDNA [33], EC-RUS [15], and Bootstrap [30], as shown in Tab. 5.

Table 4: Comparison of DTLM-DBP with previous methods for PDB 186 dataset

Method	Acc (%)	Sen (%)	Spe (%)	MCC
DNABIND [26]	67.70	66.70	68.80	0.35
DNAbinder [22]	60.80	57.00	64.50	0.22
DNA-Threader [27]	59.70	63.70	95.70	0.28
DNA-Prot [16]	61.80	68.00	53.80	0.24
DBPPred [28]	76.90	79.60	74.20	0.54
iDNA-Prot [17]	67.20	66.70	66.70	0.34
iDNA-Prot-dis [21]	80.64	80.00	80.00	0.54
PSSM-DT [8]	80.00	87.09	72.83	0.65
iDNAPro-PseAAC [20]	69.89	77.00	62.40	0.40
Mixed Feature [32]	78.95	73.68	84.21	0.58
PseDNA-Pro [23]	76.55	79.61	79.61	0.53
iDNAProt-ES [10]	80.64	81.00	80.00	0.61
PSFM-DBT [12]	80.65	90.32	70.97	0.62
Local-DPP [19]	79.00	92.00	65.60	0.63
HMMBinder [25]	69.02	61.00	76.30	0.39
DPP-PseAAC [24]	77.42	83.00	70.90	0.55
iDNA-KACC-EL [31]	79.03	94.62	63.44	0.61
DTLM-DBP (proposed)	81.18	91.43	75.00	0.64

Table 5: Comparison of DTLM-DBP with previous methods for PDNA-543 dataset

Method	Acc (%)	Sen (%)	Spe (%)	MCC
TargetDNA [33]	91.40	40.60	95.00	0.34
EC-RUS [15]	91.80	47.62	94.92	0.39
Bootstrap [30]	90.77	78.77	92.36	0.63
DTLM-DBP (proposed)	93.05	48.11	97.57	0.53

For PDNA-316 dataset, DTLM-DBP was compared with MetaDBSite [34], TargetDNA [33], EC-RUS [15], and Bootstrap [30], as shown in Tab. 6.

The results confirmed the efficacy and viability of the proposed method for different datasets.

Table 6: Comparison of DTLM-DBP with previous methods for PDNA-316 dataset

Method	Acc (%)	Sen (%)	Spe (%)	MCC
MetaDBSite [34]	77.00	77.00	77.00	0.32
TargetDNA [33]	90.99	43.02	95.00	0.37
EC-RUS [15]	91.50	49.35	95.00	0.42
Bootstrap [30]	91.03	82.47	92.34	0.67
DTLM-DBP (proposed)	93.38	55.35	97.69	0.60

4 Conclusions

The paper presented an efficient new approach for DNABPs identification based on deep transfer learning "DTLM-DBP." The protein sequences were adapted to CNN models using 1D convolutions layers, then the VGG-NET 16 pre-trained deep transfer learning models were used as a feature's extractor. Finally, the RF classifier was used for sequence features matching. DTLM-DBP was tested using different DNA proteins datasets and compared with the other published DNABPs identification methods, and it has provided a considerable improvement in the performance of DNABPs identification.

Acknowledgement: We would like to acknowledge Universiti Teknologi Malaysia for their support via the 2020-2021 Industry-International Incentive Grant which funded this publication.

Funding Statement: This paper was funded under the 2020–2021 Industry-International Incentive Grant by Universiti Teknologi Malaysia (Grant Number: Q.K130000.3043.02M12) which was granted to U. Khairuddin, F. Behrooz and R. Yusof.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Shadab, M. T. A. Khan, N. A. Neezi, S. Adilina and S. Shatabda, "DeepDBP: Deep neural networks for identification of DNA-binding proteins," *Informatics in Medicine Unlocked*, vol. 19, no. 2, pp. 100318, 2020.

[2] C. Tan, T. Wang, W. Yang and L. Deng, "PredPSD: A gradient tree boosting approach for single-stranded and double-stranded DNA binding protein prediction," *Molecules*, vol. 25, no. 1, pp. 1–16, 2020.

[3] J. Qiu, M. Bernhofer, M. Heinzinger, S. Kemper, T. Norambuena *et al.,* "ProNA2020 predicts protein-DNA, protein-RNA, and protein–protein binding proteins and residues from sequence," *Journal of Molecular Biology*, vol. 432, no. 7, pp. 2428–2443, 2020.

[4] J. Zhang, Q. Chen and B. Liu, "DeepDRBP-2L: A new genome annotation predictor for identifying DNA binding proteins and RNA binding proteins using convolutional neural network and long short-term memory," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2019.

[5] A. Trabelsi, M. Chaabane and A. Ben-Hur, "Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities," *Bioinformatics*, vol. 35, no. 14, pp. 269–277, 2019.

[6] Z. Zhao, Y. Xu and Y. Zhao, "SXGBsite: Prediction of protein-ligand binding sites using sequence information and extreme gradient boosting," *Genes*, vol. 10, pp. 9651–96519, 2019.

[7] H. J. Zhu, Z. H. You, W. L. Shi, S. K. Xu, T. H. Jiang *et al.,* "Improved prediction of protein–protein interactions using descriptors derived from PSSM via gray level co-occurrence matrix," *IEEE Access*, vol. 7, pp. 49456–49465, 2019.

[8] R. Xu, J. Zhou, H. Wang, Y. He, X. Wang *et al.,* "Identifying DNA-binding proteins by combining support vector machine and PSSM distance transformation," *BMC Systems Biology*, vol. 9, no. Suppl. 1, pp. S10, 2015.

[9] M. Waris, K. Ahmad, M. Kabirand and M. Hayat, "Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix," *Neurocomputing*, vol. 199, no. 1, pp. 154–162, 2016.

[10] S. Y. Chowdhury, S. Shatabda and A. Dehzangi, "iDNAProt-ES: Identification of DNA-binding proteins using evolutionary and structural features," *Scientific Reports*, vol. 7, pp. 1–14, 2017.

[11] R. Xuab, J. Zhoua, B. Liu, Y. Hee, Q. Zouf *et al.,* "Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach," *Journal of Biomolecular Structure and Dynamics*, vol. 33, no. 8, pp. 1720–1730, 2015.

[12] J. Zhang and B. Liu, "PSFM-DBT: Identifying DNA-binding proteins by combing position specific frequency matrix and distance-bigram transformation," *International Journal of Molecular Sciences*, vol. 18, pp. 1–16, 2017.

[13] J. Zhang, B. Gao, H. Chai, Z. Ma and G. Yang, "Identification of DNA-binding proteins using multi-features fusion and binary firefly optimization algorithm," *BMC Bioinformatics*, vol. 17, no. 1, pp. 323, 2016.

[14] X. Ma, J. Guo and X. Sun, "DNABP: Identification of DNA-binding proteins based on feature selection using a random forest and predicting binding residues," *PLOS ONE*, vol. 11, no. 12, pp. e0167345, 2016.

[15] C. Shen, Y. Ding, J. Tang, J. Song and F. Guo, "Identification of DNA-protein binding sites through multi-scale local average blocks on sequence information," *Molecules*, vol. 22, pp. 1–20, 2017.

[16] K. Kumar, G. Pugalenthi and P. N. Suganthan, "DNA-Prot: Identification of DNA binding proteins from protein sequence information using random forest," *Journal of Biomolecular Structure & Dynamics*, vol. 26, no. 6, pp. 679–686, 2009.

[17] W. Z. Lin, J. A. Fang, X. Xiao and K. C. Chou, "iDNA-Prot: Identification of DNA binding proteins using random forest with grey model," *PloS One*, vol. 6, no. 9, pp. e24756, 2011.

[18] X. Z. Fu, W. Zhu, B. Liao, L. Cai, L. Peng *et al.,* "Improved DNA-binding protein identification by incorporating evolutionary information into the Chou's PseAAC," *IEEE Access*, vol. 6, pp. 66545–66556, 2018.

[19] L. Wei, J. Tang and Q. Zou, "Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information," *Information Sciences*, vol. 384, pp. 135–144, 2017.

[20] B. Liu, S. Wang and X. Wang, "DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation," *Scientific Reports*, vol. 5, no. 1, pp. 15479, 2015.

[21] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou *et al.,* "iDNA-Prot—dis: Identifying dna-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PloS One*, vol. 9, no. 9, pp. e106691, 2014.

[22] M. Kumar, M. M. Gromiha and G. P. Raghava, "Identification of DNA-binding proteins using support vector machines and evolutionary profiles," *BMC Bioinformatics*, vol. 8, no. 1, pp. 463, 2007.

[23] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou *et al.,* "PSEDNA-pro: DNA-binding protein identification by combining chou's pseaac and physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015.

[24] M. S. Rahman, S. Shatabda, S. Saha, M. Kaykobad and M. S. Rahman, "DPP-PSEAAC: A DNA binding protein prediction model using chou's general pseaac," *Journal of Theoretical Biology*, vol. 452, no. 1, pp. 22–34, 2018.

[25] R. Zaman, S. Y. Chowdhury, M. A. Rashid, A. Sharma, A. Dehzangi *et al.,* "DNA-binding protein prediction using hmm profile-based features," *BioMed Research International*, vol. 2017, pp. 1–10, 2017.

[26] A. Szilágyi and J. Skolnick, "Efficient prediction of nucleic acid binding function from low-resolution protein structures," *Journal of Molecular Biology*, vol. 358, no. 3, pp. 922–933, 2006.

[27] M. Gao and J. Skolnick, "A threading-based method for the prediction of DNA binding proteins with application to the human genome," *PLOS Computational Biology*, vol. 5, no. 1, pp. e1000567, 2009.

[28] W. C. Lou, X. Q. Wang, F. Chen, Y. X. Chen, J. Bo *et al.,* "Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and gaussian naïve bayes," *PLos One*, vol. 9, no. 1, pp. e86703, 2014.

[29] Y. Zhang, S. Qiaoc, S. Jia, N. Hand, D. Liuc *et al.,* "Identification of DNA-protein binding sites by bootstrap multiple convolutional neural networks on sequence information," *Engineering Applications of Artificial Intelligence*, vol. 79, no. 4, pp. 58–66, 2019.

[30] Y. Zhang, S. Qiao, S. Ji and Y. Li, "DeepSite: Bidirectional LSTM and CNN models for predicting DNA-protein binding," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 4, pp. 841–851, 2020.

[31] B. Liu, S. Wang, Q. Dong, S. Li and X. Liu, "Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning," *IEEE Transactions on NanoBioscience*, vol. 15, no. 4, pp. 328–334, 2016.

[32] K. Qu, K. Han, S. Wu, G. Wang and L. Wei, "Identification of DNA-binding proteins using mixed feature representation methods," *Molecules*, vol. 22, no. 10, pp. 1602, 2017.

[33] J. Hu, Y. Li, M. Zhang, X. Yang, H. B. Shen *et al.,* "Predicting protein-DNA binding residues by weightedly combining sequence-based features and boosting multiple SVMs," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 6, pp. 1389–1398, 2017.

[34] J. Si, Z. Zhang, B. Lin and B. Huang, "MetaDBSite: A meta approach to improve protein DNA-binding sites prediction," *BMC Systems Biology*, vol. 5, no. Suppl 1, pp. S7, 2011.

[35] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Magenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*. New York, NY, USA: Curran Associates, Inc., pp. 1097–1105, 2012.

[36] S. H. Wang, S. Xie, X. Chen, D. S. Guttery, C. Tang *et al.,* "Alcoholism identification based on an AlexNet transfer learning model," *Front Psychiatry*, vol. 10, pp. 205, 2019.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, pp. 1–14, 2015.

[38] S. Khatoon, M. M. Hasan, A. Asif, M. Alshmari and Y. K. Yap, "Image-based automatic diagnostic system for tomato plants using deep learning," *Computers, Materials & Continua*, vol. 67, no. 1, pp. 595–612, 2021.

[39] T. Y. Heo, K. M. Kim, H. K. Min, S. M. Gu, J. H. Kim *et al.,* "Development of a deep-learning-based artificial intelligence tool for differential diagnosis between dry and neovascular age-related macular degeneration," *Diagnostics*, vol. 10, no. 5, pp. 1–10, 2020.

[40] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Leaming*, vol. 20, pp. 273–297, 1995.

[41] H. Kasban, "Fingerprint's verification based on their spectrum," *Neurocomputing*, vol. 171, no. 1, pp. 910–920, 2016.

[42] A. Abozaid, A. Haggag, H. Kasban and M. Eltokhy, "Multimodal biometric scheme for human authentication technique based on voice and face recognition fusion," *Multimedia Tools and Applications*, vol. 78, no. 12, pp. 16345–16361, 2019.

[43] J. S. H. Al-bayatiand and B. B. Üstündağ, "Fused and modified evolutionary optimization of multiple intelligent systems using ANN, SVM approaches," *Computers, Materials & Continua*, vol. 66, no. 2, pp. 1479–1496, 2021.

[44] V. N. Vapnik, *Statistical Learning Theory*. New York, USA: Wiley, 1998.

[45] T. K. Ho, "Random decision forests," in *Proc. ICDAR*, NW Washington, DC, USA, pp. 278–282, 1995.

[46] P. Calhoun, M. J. Hallett, X. Su, G. Cafri, R. A. Levine *et al.,* "Random forest with acceptance-rejection trees," *Computational Statistics*, vol. 35, no. 3, pp. 983–999, 2020.