Journal of Theoretical and Applied Information Technology
31st August 2022. Vol.100. No 16
© 2022 Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A NOVEL SEQUENCE-BASED NEGATIVE SAMPLING APPROACH FOR IMPROVING PROTEIN-PROTEIN INTERACTIONS PREDICTION USING MACHINE LEARNING TECHNIQUES

M. SAYED BARKAT[1], SHERIN M. MOUSSA [2], NAGWA L. BADR [2]

[1,2]Department of Information Systems, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt

E-mail: [1]Mohamed_barkat@cis.asu.edu.eg, [2]sherinmoussa@cis.asu.edu.eg, [2]Nagwabadr@cis.asu.edu.eg

ABSTRACT

Protein–protein interactions (PPIs) have been involved in numerous diseases’ progression in drug discovery. Although PPIs prediction is a crucial and well-studied task in bioinformatics, they still lack thorough investigations for several proteins. The cost of understanding PPIs and identifying protein–protein non-interactions (PPNIs) using sequence alignment make the current computational methods inefficient, so identifying PPNIs without applying sequence alignment has become a necessity. In this research, a machine learning approach is proposed for PPIs prediction based on protein sequence information, in which we introduced “Features-based Negative Generation” which is a novel approach for identifying PPNIs samples. This method measures sequence features' similarity without alignment for an affordable computational feasibility. After PPNIs identification the Conjoint Triad (COT) and Epitopes are used for features extraction and results of both are compared to achieve higher accuracy with less time consumption. Five machine learning techniques were investigated to learn from the interacting pairs sequence, obtaining PPI features. Support vector machine (SVM) with polynomial and RBF kernel functions, Linear SVM, Tree Model (TM) and Linear Model, and the (TM) achieved the best result with an accuracy of 97.8%. The experimentation of PPIs prediction using generated negative dataset and COT using 343 features achieved an accuracy of 97.8%, versus 93% using random negative dataset using COT also. Applying Epitopes with our PPNIs dataset using 21 features achieved an accuracy of 94.5% versus 92.5% with random negative dataset, which indicates that identified PPNIs datasets are clearer, less noise and prediction of PPI using identified PPNIs is more accurate. We compared PPI prediction accuracy using identified PPNIs which extracted using our method with that obtained by other methods in the literature, and we found improvement in our favor of between 2 and 7%. Considering Epitopes for features extraction is faster than COT by an average of 83%.

Keywords: *Protein-Protein Interaction, Protein–Protein Negative Interactions, Machine Learning, Biological Pathways, Drug Discovery, Ppnis Sampling, Epitopes, Conjoint Triad Method.*

1. INTRODUCTION

Protein-protein interactions (PPIs) have encountered a great interest in biology, as they have a critical role in regulating roughly all cellular biological processes, including DNA metabolic reactions & replication, cellular organization and immune response [1,2] . PPI dysfunctions have been implicated in various diseases, where the understanding of PPI mechanisms is very useful in the disease biology research [3]. Among the molecular networks, Protein-Protein Interaction Networks (PINs) have become effective platforms for uncovering the molecular mechanisms of diseases and drug discovery [4]. Thus, PPIs analysis has shed the light on drug target detection methodologies and aided in therapy design [5]. Given the PPIs importance, the prediction of PPIs is vital, as it helps researchers make reasoning and conclusions of PPI outcomes, in which PPIs can be studied by intracellular localization, phylogenetic profiling and post-translational modifications [6]. Different machine learning approaches have been considered for PPIs prediction, in which the model needs to learn how to differentiate between positive and negative classes, having insufficient accurate information to build such a model that differentiates between them [4,7,8]. To collect PPI interactions, great-scale and high-throughput experimental approaches, such as yeast two-hybrid [9], tandem affinity purification [10] and mass spectrometry [11], were used. However, researchers frequently pointed out that these methods suffer various drawbacks, including the cost, time-consumption, and inaccurate results. Consequently, different computational methods were built to predict PPI interactions, while avoiding those drawbacks [12].

In this paper, a machine learning approach for PPIs prediction within the human body is introduced based on protein sequence information, in which the "Features-based Negative Generation" method is proposed to identify PPNIs datasets from large PPI datasets using sequence features' similarity and without using alignment This is based on several studies that have worked on the same goal such as in [23] and in [4,7] using sequence similarity but with alignment. Five machine learning techniques were evaluated, such as support vector machine (SVM) with polynomial and RBF kernel functions, Linear SVM, Tree Model and Linear Model for PPI data classifications, considering two features extraction methods; Conjoint Triad (COT) and Epitopes. Thus, the main contributions of this study can be summarized as follows:

1. A novel method is proposed to generate negative datasets from large PPI datasets with neither random sampling nor alignment, in which the similarity among proteins is measured between their features for negative sampling, saving the time of alignment and comparing results of our method with results of other methods in literature to highlight the accuracy of our method for identifying PPNIs.
2. Two features extraction methods were investigated to achieve optimum results with fewer features and less time consumption and this the first time for using Epitopes in PPI prediction.
3. False negatives are reduced dramatically. Our machine learning-based prediction approach is efficient and effective in predicting protein-protein interactions.

2. RALTED WORK

The most popular machine learning techniques in PPIs prediction are the Support Vector Machines (SVMs) and Random Forest (RF) classifiers [4,7,8,13–16]. Variant researches attempted to use SVMs with extracted features from protein sequences. In [4], authors tried to predict PPIs based on the features of proteins sequences, where the conjoint thread method (COT) was used for features extraction with SVM. This achieved an accuracy level of 83.90% on more than 16,000 diverse PPI pairs. Yet, the considered number of features was huge, where each protein and each interaction were represented in (7*49) and (7*98) matrices respectively. Due to the huge number of extracted features, COT was time-consuming, especially as the volumes of data increase. Besides, only one dataset was investigated (HPRD 2005), which was insufficient for the results validity. The used PPNIs were chosen randomly, which caused noise in data. In [17,18], negative datasets were selected randomly, achieving good results as shown in table 1, but random selection caused considerable noise in data as well.

Several studies have been conducted to minimize the noise in negative datasets resultant from random selection by developing methods to generate negative datasets. In [7], a machine learning framework was presented to predict PPIs of viral proteins with human proteins using a negative sampling method to generate negative samples of a novel virus with a host, due to the lack of viral protein interactions data. The PPNIs were selected after reducing them by restricting the PPNIs based on the similarity between the viral proteins' sequences. This reduced the noise in the generated PPNIs, achieving an accuracy level of 81% on a complete simulation, and up to 86% on a partial simulation. COT was used for features extraction, which was time-consuming. In addition, the training was conducted over 5,753 interactions among 2,357 human proteins and 453 viral proteins, which was considered a very small number of proteins for alignment. Using alignment with huge datasets in negative sampling to measure similarity between viral protein sequences and generate less noisy PPNIs would be very ineffective, as the complexity of Needleman-Wunsch alignment ranges from (O(mn)+O(n)) to O(mn) [19,20].

Some studies tried to identify remote evolutionary relations among proteins as in [15], which obtained pairwise protein descriptors. This codification allowed generating up to 13,248 features then reduced to 322 using Weka. The features were extracted from 4,327 protein pairs, divided into 1,922 interacting pairs and 2,405 non-interacting pairs. Their method depended on protein domains, identifying active EPI-X4 derivatives based on the predicted interaction with CXCR4 fragments by an accuracy up to 71%. In [16], a technique was developed to predict human-virus

PPIs using a computational approach. Positive and negative data samples were constructed using human-virus protein interactions data from SwissProt database [21]. A corpus of sequence information from such protein data was created to train a doc2vec model in order to extract protein sequence features, predicting PPIs using random forest by an accuracy up to 93.2%. In [22], another approach was developed for PPI prediction using the NIP-SS and NIP-RW methods to obtain a negative dataset, with an accuracy of 86.17% and 86.44% for NIP-SS and NIP-RW respectively with H. sapiens dataset. However, both methods depended on the use of Needleman-Wunsch alignment, which was dramatically time consuming to measure similarity among huge number of proteins.

Other studies in PPI prediction tried to develop alternative methods for negative datasets generation without using alignment. In [23], authors worked on predicting interactions in yeast using a variety of data sources. 10,517 interactions among 4,233 yeast proteins were used, in which gene ontology (GO) similarity between proteins was considered to generate negative dataset. However, this led to biased estimates of prediction accuracy. A negative dataset was generated in [24] using the RandomPairs and RecombinePairs methods, achieving an accuracy of 81% and 52% respectively using 78,000 PPIs among more than 27,000 proteins. Yet, both negative generation methods depended on random selection in their core, causing noise in the generated dataset and low-quality data. Table 1 provides a comparative study between a set of technical approaches applied in several scientific researches and our research in terms of datasets used in the experiments and their properties, features extraction methods, machine learning classification methods, negative generation method which in turn also indicates whether it depends on alignment or not and it also highlights the superiority of our research in the accuracy of PPI prediction.

Table 1: A comparative summary for the PPI prediction approaches considering negative sampling.

Refs.	Data set	Classifier	Feature extraction method	Use Alignment?	Negative generation method	Accuracy
[7]	Virus Mentha (5753 interactions between 2,357 human proteins and 453 viral proteins)	SVM	COT	Yes	Dissimilarity-based negative sampling	86%
[22]	H. sapiens (1,412 interactions)	Deep neural networks	COT	Yes	NIP-SS	86.17% ± 0.93%
					NIP-RW	86.44% ± 0.59%
[4]	HPRD database (version 2005)	SVM	COT	No	Random sampling	83.90 ± 1.29
[18]	HIPPIE [27]	Random Forest (RF)	PSSM			85%
[17]	DIP database (5081 H. sapiens)	Long Short-Time Memory (LSTM)	node2vec model			83%
[23]	BIND, MIPS and DIP [28-30]	SVM	GO features	No	based on similarity of GO annotations	ROC: 0.874
[24]	DIP database (5346H. sapiens)	Ensemble Classifier (LibD3C), with clustering & dynamic selection strategy	COT	No	RandomPairs	81.9%
					RecombinePairs	52.5%
[12]	HPRD database	SVM	COT	No	Random sampling	93,45%
[25]	1412 interactions (H. sapiens)	LightGBM-PPI	PseAAC	No		94.83%
[26]	45,856 banana protein sequences with 14459 Foc4 protein sequences	LSTM	COT	No		94%

Table 1 Representation for the main methods which using ML techniques in PPI that tried to find ways for negative sampling generation. The random selection method is widely used, although it causes noise in the negative data. RandomPairs and RecombinePairs methods were applied in some studies for negative interactions dataset generation on sequence features for PPI prediction without

using alignment. Hence, such methods would be investigated for our comparative experimentation to compare results using the same datasets. We will not consider the comparison of our results with methods using alignment, due to the difficulty of applying alignment on large volumes of data. In addition, we consider only the methods that depend on sequence features [23].

3. METHODS

In this section we describe our methodology for predicting interactions between proteins, where we first began by explaining the hypotheses on which we based our theory in determining PPNIs ,and then we explained the “Features-based Negative Generation” is presented as a novel negative sampling approach based on sequence data to minimize noise prediction resulting from random negative samples, then we determined the features of the proteins that we relied on to distinguish between PPIs and PPNIs and the used features extraction methods to ensure high levels of accuracy and reduced time consumption .Finally, we explain the machine learning methods that used to learn from the sequences of interacting pairs for the PPIs prediction process, why they were used and how they are applied and this was following the methodology of research that works for the same purpose[7].

3.1 Hypotheses

First hypothesis: Features-based learning model. We assume that if we train a classification model using a large number of positive and negative interactions between different human proteins, this will make it eligible for proteins interactions classification. A classifier should extract the features associated with the proteins whether interacting or not. Based on this hypothesis, the prediction of PPI of any human protein would be possible, as the model can predict whether a protein has any interactions with the trained proteins.

Second hypothesis: Similarity in sequence features, similarity in interactions. We assume that the proteins with high sequence features similarity can have a large number of similar interactions theoretically. We relied on this hypothesis to propose a method that deduces negative interactions to train the model, where using negative random sampling possibly leads to false negative interactions [7]. We anticipate that strong sequence features similarity between tested and training proteins enhances the classifier accuracy. As per these hypotheses, we propose a method for negative dataset generation.

3.2 The Proposed Features-based Negative Generation Approach

In this section we explained how to identify PPNIs using our Negative generation approach

Previously, it was relied upon random sampling in order to get negative interactions. In the random sampling method, for protein X, all other human proteins in the dataset rather than its interactions are considered as negative interactions with X, and then these proteins are chosen randomly. According to Hypothesis 2, random sampling production of many incorrect negative samples is expected, which leads to decreasing the prediction sensitivity and accuracy, urging to find a new method to solve this problem.

Proposed sampling. Our proposed method for negative generation aims to generate more accurate negative interactions and decrease the expected false negative samples compared to random sampling. According to our hypotheses, if two proteins are similar in sequence features, a protein that interacts with one of them cannot be considered a negative sample for the other. Thus, we introduce the Features-based Negative Generation algorithm, which calculates the Euclidean distance among all vectors of proteins features. The distances are then normalized and considered as dissimilarity distances matrix for all proteins, instead of using alignment to meager similarity between proteins that needs a lot of time if the number of proteins is huge. The proteins having the lowest positive interactions are selected (negative proteins). Based on the application of dissimilarity average threshold and dissimilarity distances matrix, major false negative interactions are excluded, then random sampling is applied over the remaining negative interactions as shown in Algorithm1.

Dissimilarity average threshold (DAT). Negative interactions are generated based on the negative proteins having the highest negative interactions. In our proposed method, we find proteins having dissimilarity with negative proteins more than (DAT), which is the average of dissimilarity distances among all vectors of proteins features, considering their interactions as negative interactions for negative proteins. If two proteins are very dissimilar in sequence, the interactions of each protein differ from the other [1]. The similarity is calculated based on the features extracted from the sequence-based on epitopes extraction from protein sequences , allowing the learning model for PPI prediction to be constructed.

3.3 Feature Extraction

Various feature extraction approaches have been considered in PPIs in order to predict PPIs. These approaches include sequence–based [31,32] , structure-based [33,34]and domain-based approaches [34,35] . Sequence-based approaches are the most commonly used for PPIs prediction[12], in which transferring known protein interactions with unknown interactions could be done based on sequence similarity due to its data availability. It can also be done based on the structure similarity [36] or sharing of interaction interfaces [7]. We used two methods for feature extraction: Conjoint Triad Method and it is commonly used in PPI prediction or with protein generally as represented in table 1 [4,37,38] and epitopes features vector extraction from protein sequences [39].

Conjoint Triad Method clusters the 20 amino acids into seven groups based on the similarities of their dipoles and volumes of the side chains then similar amino acids are clustered in the same class. Each frequent 3-mers (three continuous amino acids) are regarded as a unit [4]. So we can obtain 7*7*7=343 tried types and the frequency of occurrence of each triad $f_i (i = 1,2,\dots,343)$ is calculated. The 343-dimensional feature vector is then calculated using equation (1) [25]. Finally, a normalized (7*49) feature vector is then generated for each protein independently. For each positive and negative dataset, the normalized feature vectors of each pair are concatenated into one feature vector to represent the interaction [7].

$$d_i = \frac{f_i - \min\{f_1, f_2, \dots, f_{343}\}}{\max\{f_1, f_2, \dots, f_{343}\}}, i = 1,2,\dots,343 \qquad (1)$$

The second method for feature extraction is based on epitopes extraction from protein sequences . An antibody, is a protein produced by the immune system that binds with high specificity to an antigen, which in turn binds only to a segment of the protein known as an epitope . Sequences are first converted into fixed length feature vectors to represent the structural and physicochemical properties of the peptides [39]. It clusters the 20 amino acids into three groups based on hydrophobicity, defining the three descriptors that describe the global composition of the epitopes as follows:

(1) Composition (C), a vector consisting of 3 real numbers, each number corresponds to the fraction of amino acids for each group.
(2) Transition (T), a vector of 3 real numbers that characterize the fraction of amino acids frequency from a group, followed and preceded by amino acids from another group.
(3) Distribution (D), a vector consisting of 15 real numbers, each 5 real numbers represent the first, 25%, 50%, 75% and 100% of amino acids for each group in the sequence [39].

Next, 21 feature vectors are generated for each protein independently. For each positive and negative dataset, the feature vectors of each pair are concatenated into one feature vector to represent the interaction.

Algorithm1: Features based negative generation approach

Input: Negative proteins vector (contains proteins having the lowest positive interactions in positive interaction database), distancesP matrix (contains distances between the negative protein and all proteins.) PositiveSetList (dictionary that contains all proteins and their interactions)
Output: Negative interactions in a container array
Initialize NegativeSetList as empty container array for saving negative interactions

```
Begin
Initialize proteins list as empty list // contains header row of distancesP
For each I in Negative proteins // for all negative proteins
    Initialize PsIncluded as empty list //to save all proteins that have dissimilarity distances with I < DAT.
    For each J in proteins_List
        If (dissimilarity distance between I and J < DAT = True)
            Add J for PsIncluded //get dissimilarity distance from distancesP and check if distance < DAT.
        End if
    End for
    Initialize HsOFF as empty list // all proteins that make interaction with proteins in PsIncluded
    For each K in PsIncluded
        Initialize Interactions as empty list//proteins that make interaction with proteinsK
        Interactions =PositiveSetList [K].
        Add Interactions to HsOFF.
    End for
    Initialize HsofP as empty list // contains all proteins that make interaction with I (positive interactions)
    HsofP = PositiveSetList [I].
    Remove HsofP from HsOFF
    Initialize Number Pos_count = length of HsofP list.
    Initialize Neg_count = Pos_count.
    Initialize a list NegSet_P as empty list // to save negative interactions of the negative protein
    If (Neg_count < length (HsOFF) and Neg_count > 0)
        NegSet_P=random number of HsOFF items equal Neg_count
    Else if (Neg_count = length (HsOFF))
        Set NegSet_P = HsOFF
    Else
        NegSet_P = empty list
    End if
    NegativeSetList (I) = NegSet_P
End for
Return NegativeSetList
END
```

3.4 Classification and Learning Model

In this section, the model is trained, validated and tested using the testing set, different machine learning techniques are used for generating machine learning model which is trained for PPI prediction, such as SVM and its kernel functions, where SVMs are used for their classification power and ability to stand high noise [31,40]. We used kernel functions such as:

Linear Kernel: $F(\,x.xj) = sum(\,x.xj)$ (2)
where x, xj represent the data to classify.

Polynomial Kernel: $F\ (x, xj) = (x.xj\ +\ 1)\hat{}d$ (3)
where '.' is the dot product of both values, d denotes the degree, F (x, xj) is the decision boundary to separate the given classes. RBF Kernel:
$F\ (x, xj) = exp\left(-gamma * \ ||x\ -\ xj||\hat{}2\right)$ (4)
where gamma defines how far the influence of a single training example reaches in a range from 0 to 1, having low values considered as 'far' and high values considered as 'close'. The gamma value ranges from 0 to 1, in which the most preferred value is 0.1 [41].

The Tree Model is used as in big databases, in which decision trees have been widely employed for both exploratory data analysis and predictive modelling applications due to these properties, as well as their intuitive interpretation [42,43]. The Linear model [40] is also used to check the nature of our dataset. As for the data partitioning and preparation, the searches were conducted at each value of the DAT in the range of [0,1], with a 0.1 step. The positive and generated negative datasets are divided into 80% as a training data and 20% as a testing data for the classification and predictive models to find the highest accuracy and sensitivity for performance evaluation.

4. EXPERIMENTAL EVALUATION

4.1 Dataset

PPIs prediction is a binary classification problem, which requires to build positive and negative datasets for experimentation. Human-Human PPIs were obtained from two different datasets as positive dataset. Firstly, the Human Protein References Database (HPRD), V2007 . This HPRD version contains 36,630 unique PPIs, among 9,776 human proteins. This dataset was used to compare our negative sampling method in negative dataset construction with the negative random selection method, as this dataset was used with random selection method in [4]. Secondly, DIP [30] that has collected more than 78,000 PPIs, among more than 27,000 proteins, where 5,346 PPIs are obtained from the DIP as human PPIs until 2013. This dataset was used to compare our negative sampling method with RandomPairs and RecombinePairs negative sampling methods, as this dataset was used with both of them in [24] for the first time in negative dataset generation.

Negative dataset: Five PPNIs datasets were used: the first dataset was obtained from [44], consisting of 36,630 PPNIs among proteins. This dataset was chosen because it contains PPNIs among a large number of proteins mentioned in (HPRD), V2007. Therefore, it can be considered as a negative random selection dataset (Downloaded Negative Data). The second dataset was generated using our Features-based Negative Generation method on (HPRD), V2007 (HPRD generated negative dataset) [45]. The Third and fourth datasets were generated using RandomPairs and RecombinePairs negative sampling methods on DIP dataset (RandomPairs negative dataset), (RecombinePairs negative dataset) respectively. The fifth dataset was generated using our proposed Features-based Negative Generation method on DIP dataset (DIP generated negative dataset).

A total of 36,630 PPNIs were selected in (Downloaded Negative Data) as the first negative dataset to ensure that 1:1 is the ratio of Positive to Negative data. Generated negative dataset (HPRD generated negative dataset) also contains 36,630 PPNIs to avoid differences in the quantity of negative datasets used in comparing the two types of data. The third, fourth and fifth contains 5,346 PPNIs to avoid differences in the quantity of negative datasets used in comparing the three types of data. The first and second negative datasets are used to compare the generated negative data with our method and negative random selection data. The third, fourth and fifth datasets are used to compare the generated negative data with our method and the generated data using RandomPairs and RecombinePairs. Each protein is given by its UniProtKB identifier, in which the sequences of proteins were retrieved from [46]. Along with the positive interactions, there were PPNIs, divided into training and testing pairs to build the proposed model.

4.2 Experimental Methodology

Each protein is represented as a vector of features depending on the applied features extraction method. A negative dataset was generated from both datasets using our proposed Features-based Negative Generation method. The negative interactions are generated based on the negative proteins having the highest negative interactions more than DAT. We used different thresholds as 0.1, 0.2, 0.3, 0.4, 0.5. The negative data were compared with the positive dataset to find the similarity ratio between them and whether the size of negative data converges with the size of positive data.

Based on our experiments, DAT =0.2 was found to be the best threshold, as the size of negative data was conjugated with the size of positive data and the similarity ratio was the lowest among all trials. Thus, the negative data generated based on DAT=0.2 were used in our experiments to apply the learning model for PPIs prediction and generated negative dataset from (HPRD), V2007 were uploaded in [45]. Finally, different machine learning algorithms are applied on the positive dataset, as well as the downloaded and generated negative datasets from our method and other methods (RandomPairs and RecombinePairs.) to build a learning model for the PPI prediction and compare results among different experiments. Experiments were conducted on a machine with an Intel(R) Core(TM) i7-7500U CPU 2.70GHz, 2.90GHz with 8GB RAM. The scikit-learn library version 0.24 was used to support machine learning in the Matlab language version MATLAB 2015b.

4.3 Experimentation and Results

This section demonstrates the experimental results of applying five machine learning techniques as follows: SVM with polynomial function (SVM(P)), SVM with RBF (RPF), Linear SVM (SVM), Tree Model (TM) and Linear Model (LM) in each experiment separately. The Discussion on Performance Evaluation section compares the results of these different experiments to highlight the positive effect of the proposed negative sampling method on PPI prediction accuracy.

Accuracy measures. To evaluate the performance of the proposed framework, we use the following measures: the total number of True Negative (TN), True Positive (TP), False Positive (FP), and False Negative (FN) incidents.

In this study, we employed four metrics to measure the classification performance of our proposed approach as follows: (1) The overall classification accuracy (ACC) as the proportion of correctly predicted interacting and non-interacting pairs to the total number of pairs [47], (2) Recall (REC) as the proportion of correctly predicted interacting pairs to the total number of correctly predicted interacting and incorrectly predicted non-interacting pairs [48].(3) Precision (PRE) as the ratio of correctly predicted interacting pairings to the sum of correctly and incorrectly predicted interacting pairs [49]. (4) F1-score as the twice the ratio between the product of PRE and REC and the product of their sum [50]

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$REC = \frac{TP}{TP+FN} \quad (6)$$

$$PRE = \frac{TP}{TP+FP} \quad (7)$$

$$F1-score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (8)$$

The proposed approach was also evaluated using the area under receiver operator characteristic curves (AUC). The receiver operator characteristic (ROC) charts are very useful and effective curves that are frequently used to analyze and show the performance of classifiers. The sensitivity and specificity of the area under the ROC curve indicator are calculated, and it is widely regarded as one of the top performance indicators [51,52].

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (9)$$

1) Experiments of Downloaded Negative Data.

In these experiments, we investigate using COT and Epitopes as two features extraction methods, while applying the considered machine learning methods on the (HPRD), V2007 positive and downloaded negative datasets to study the effect of the extracted features on the PPI prediction model on both datasets, and then compare the results of downloaded negative dataset experiments with those of our generated negative dataset experiments.

Exp1: Applying machine learning methods on downloaded negative dataset using COT

In this experiment (HPRD), V2007 positive and downloaded negative datasets are converted to COT vectors and concatenated into one matrix (7*98) as an interaction to test the results using 686 features for each interaction with the mentioned datasets in order to compare results with the results of other experiments. As shown in Figure 1, SVM(P) and TM have the best accuracy, which was 93% and it also noted TM and SVM(P) outperforms all other classifiers in all evaluation metrics but SVM(P) shows between precision outperforming TM by 1% and they were almost equal in other evaluation metrics. As the best and the lowest classifiers, SVM(P) surpasses (LM) in accuracy, recall, precision, F1-score and AUC by 18.69%, 23.73%, 16.32%, 20.31% and 18.68% respectively. During the experiment, it was noted that SVM, SVM(P) and RBF consumed an average of two hours more than TM and LM, which means that 686 features for each interaction were used making SVM slow compared to TM. We also conclude from the high results SVM(P) and TM compared to the rest of the classifiers that the features extracted by COT were enough for them to make classification process more accurate than the rest of the classifiers. In general, all methods consumed an average of 1-3 hours because of the number of features in the huge size of datasets.

Exp2: Applying machine learning methods on downloaded negative dataset using Epitopes

In this experiment ML methods were applied on the (HPRD), V2007 positive and downloaded negative datasets after converting them to Epitopes vectors and concatenated into one matrix (1*42) as an interaction to test the results using 42 features for each interaction. As shown in Figure 2, TM is the best in all evaluation metrics as 92.5%, 92.46% ,92.54% for accuracy, F1-score and AUC respectively. The recall of TM outperforms SVM, SVM(P), RBF and LM by 16.9%, 20.63%, 19.12%, 20.72% respectively. Precision is superior at TM than SVM, SVM(P), RBF and LM by 23.43%, 26.06%, 25.25%, 26.09% respectively. We also noted that this experiment consumed around 30 minutes although the dataset was huge, but the number of extracted features was small. Besides, there is a great convergence in the accuracy ratios between the used machine learning methods, with an average accuracy between 68% and 71%, except for TM that was 92.5%, which demonstrates that TM is more in line with the data nature than the other ML classifiers. In addition, it can be concluded that the features extracted by Epitopes are sufficient for TM to make the process of classification more accurate, unlike the rest of the classifiers. TM in comparison with the other classifiers, it took 25% of their time. Our investigation also reveals that the data nature is overlapping, since the LM approach was the least accurate.

2) Experiments of Downloaded Negative Data.

In these experiments, we investigate using COT and Epitopes as two features extraction methods while applying the considered machine learning methods on the (HPRD), V2007 positive datasets and the generated negative dataset using the proposed Features-based Negative Generation approach on HPRD dataset to study the effect of extracted features on the PPI prediction model on both datasets, comparing the results of our generated negative dataset experiments with those of the downloaded negative dataset in the Discussion on Performance Evaluation section.

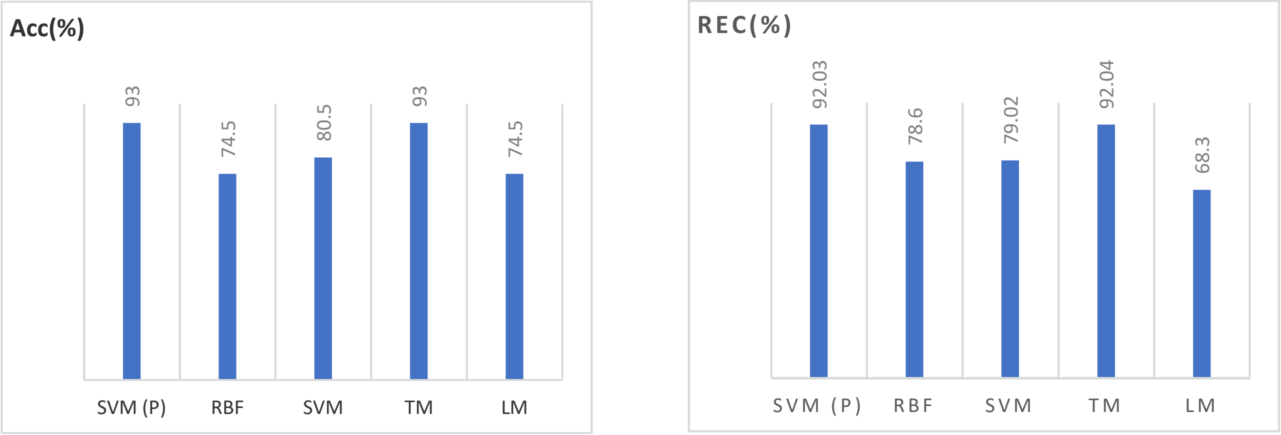
Exp3: Applying machine learning methods on the (HPRD generated negative dataset) using Epitopes.

The machine learning methods were applied on the (HPRD), V2007 positive and generated negative datasets after converting them to Epitopes vectors and concatenated into one matrix (1*42) as an interaction to test the results using 42 features for each interaction via the negative dataset generated based on the proposed Features-based Negative Generation method. As shown in Figure 3, TM is the best in all evaluation metrics as 94.5%, 95% ,94.64% for accuracy, F1-score and AUC respectively, when the negative dataset was changed to (HPRD generated negative dataset) and still using small number of features. All used classifiers achieved good recall percentages, which were 84.06, 83.03, 79.32, 94.01, and 79.33 for SVM(P), RBF, SVM, TM and LM respectively. Precision is superior at TM than SVM, SVM(P), RBF and LM by 36.07%, 29.67%, 31.43% and 36.06% respectively.

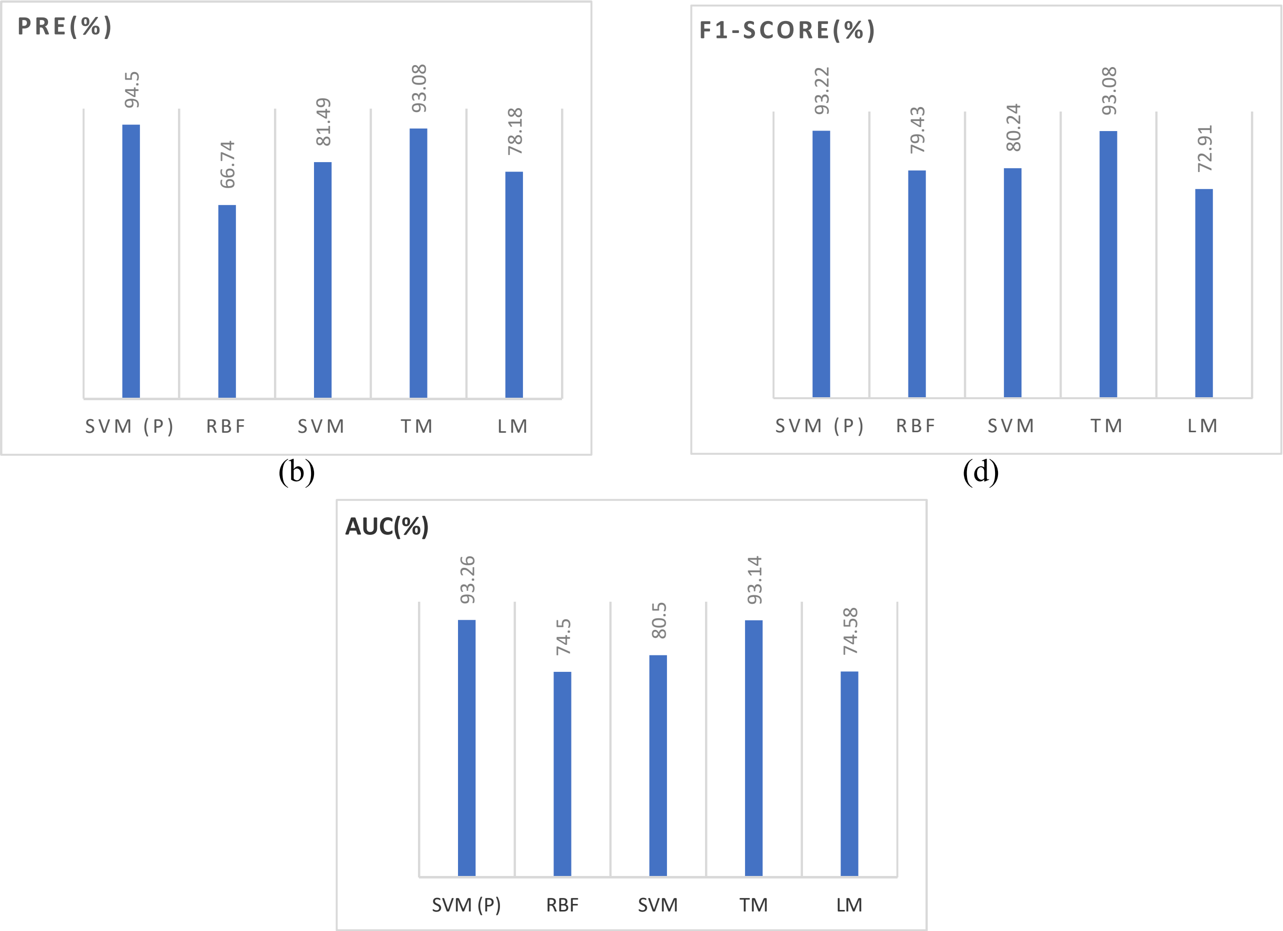
It was noted that this experiment took around 30 minutes, because of the small number of extracted features. It was observed a decrease in the results of most of the methods except for TM, which proves that the features extracted by Epitopes are sufficient for TM to make the process of classification more accurate than the rest of the classifiers by comparing the results of TM in this experiment with TM results in Exp2, whereas both of them used Epitopes for features extraction.

Exp4: Applying machine learning methods on the (HPRD generated negative dataset) using COT.

The machine learning methods were applied on the (HPRD), V2007 positive and generated negative datasets after converting them to COT vectors and concatenated into one matrix (7*98) as an interaction to test the results using 686 features for each interaction using the negative dataset generated based on the proposed Features-based Negative Generation method. As presented in Figure 4, TM has the best accuracy, which outperforms SVM(P), RBF, SVM and LM by 2%, 21.53%, 25.84% and 27.41% respectively. SVM(P) consumed an extra average of 1.5 hours compared to TM, as 686 features were used with the huge dataset size. The recall of TM outperforms SVM(P), RBF, SVM, and LM by 4.13%, 0.28%, 24.63%, 24.36% respectively. It is clear that the difference in recall between TM and RBF almost does not exist. Precision is superior at TM than SVM(P), RBF, SVM and LM by 1.44%, 33.83%, 20.04% and 29% respectively. TM achieved the best F1-score and AUC, which were 97.88 and 97.89 respectively. It was also noted that SVM(P) results are similar to TM in all evaluation metrics and both have the highest results, this is what was observed in Exp1 as well, which proves that the features extracted by COT were enough for them to make classification process more accurate than the rest of the classifiers. When the negative dataset was changed to (HPRD generated negative dataset). while using the same feature extraction method (COT) that was used in Exp1, the accuracy has increased by 4%, achieving the best accuracy of all experiments using (HPRD dataset), whereas the accuracy of TM was 97.8%.



(a) (c)

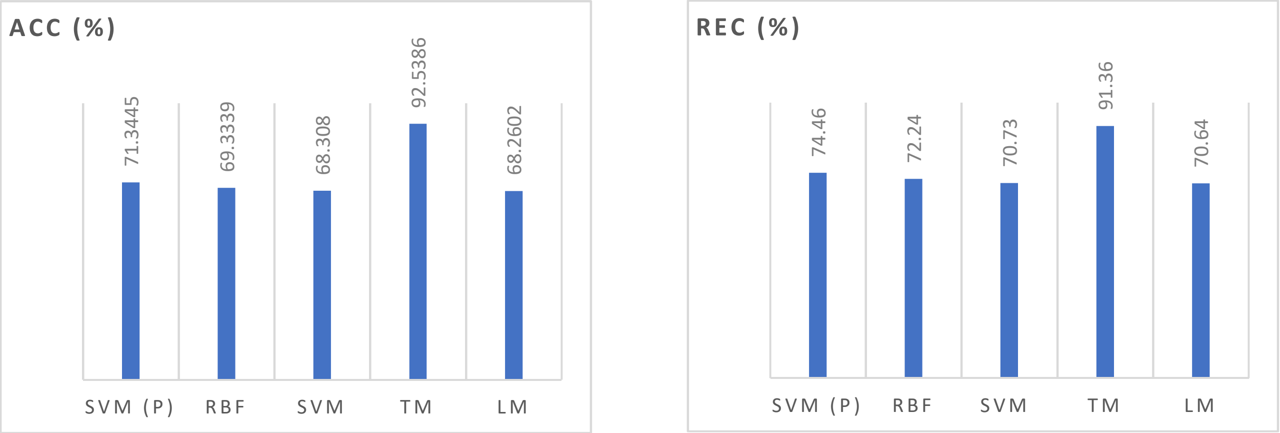


(b)

(d)

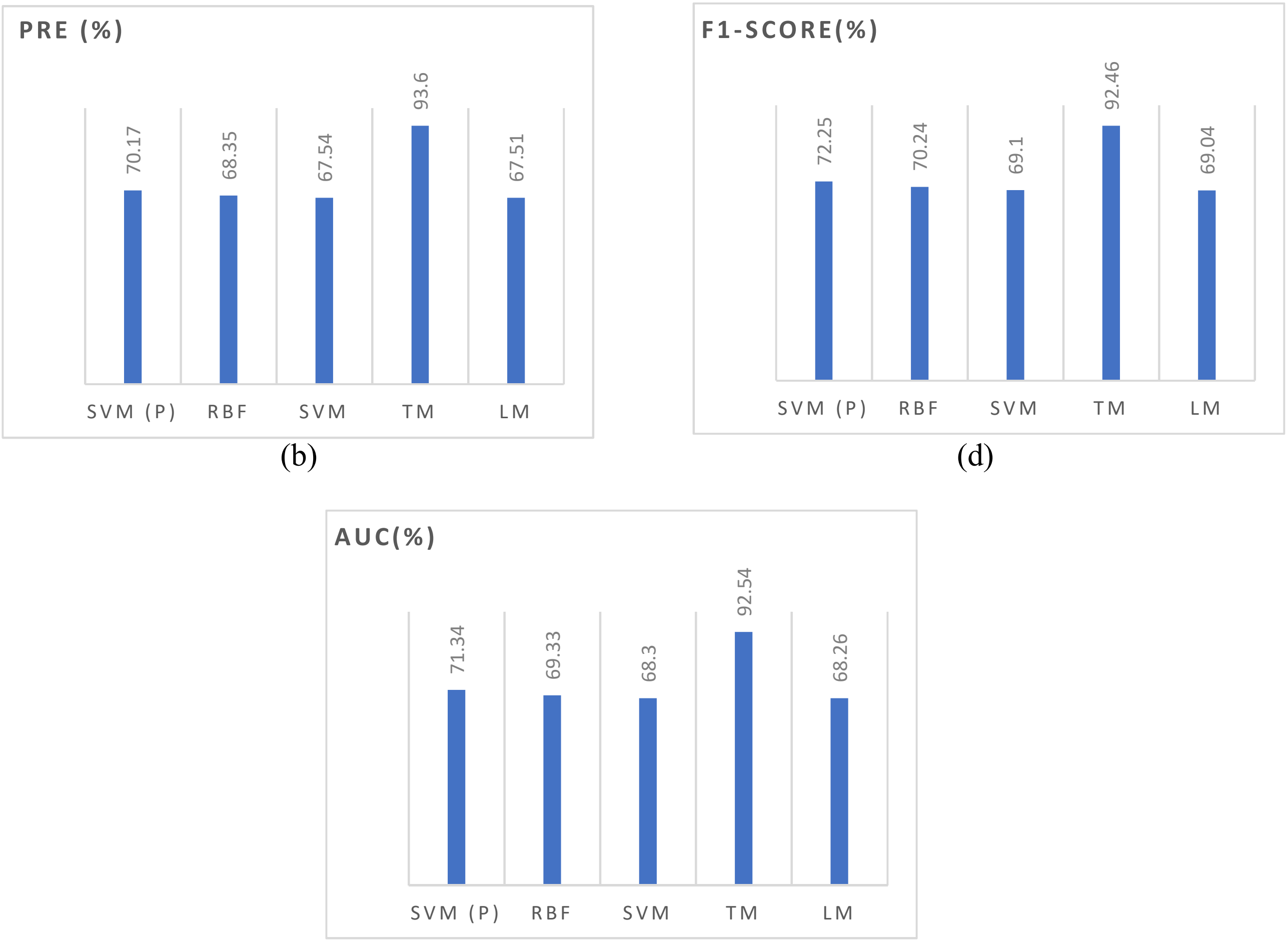
(e)

FIGURE 1. THE COMPARISON OF SVM(P), RBF, SVM, TM, AND LM USING COT ON THE DOWNLOADED NEGATIVE DATASET



(a)

(c)

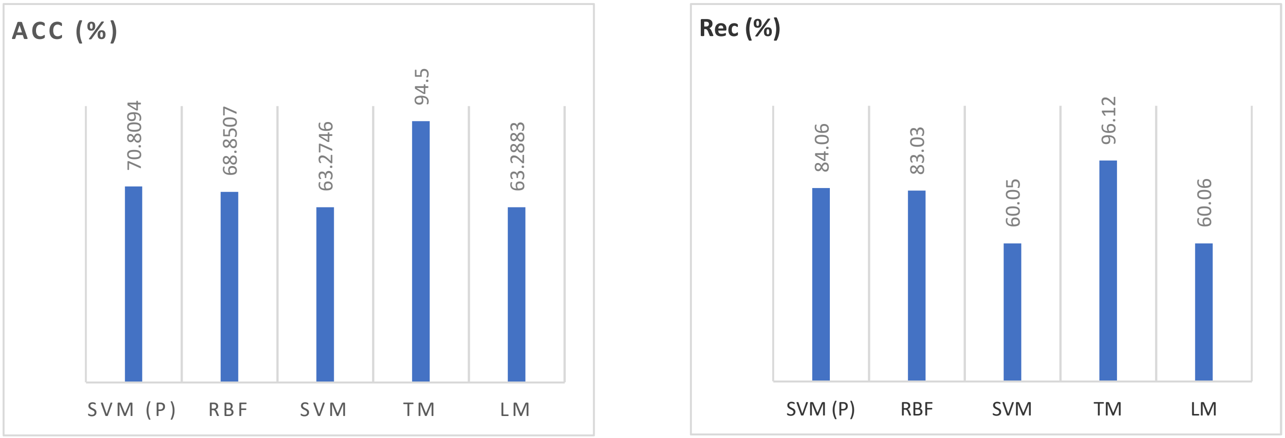


(b)

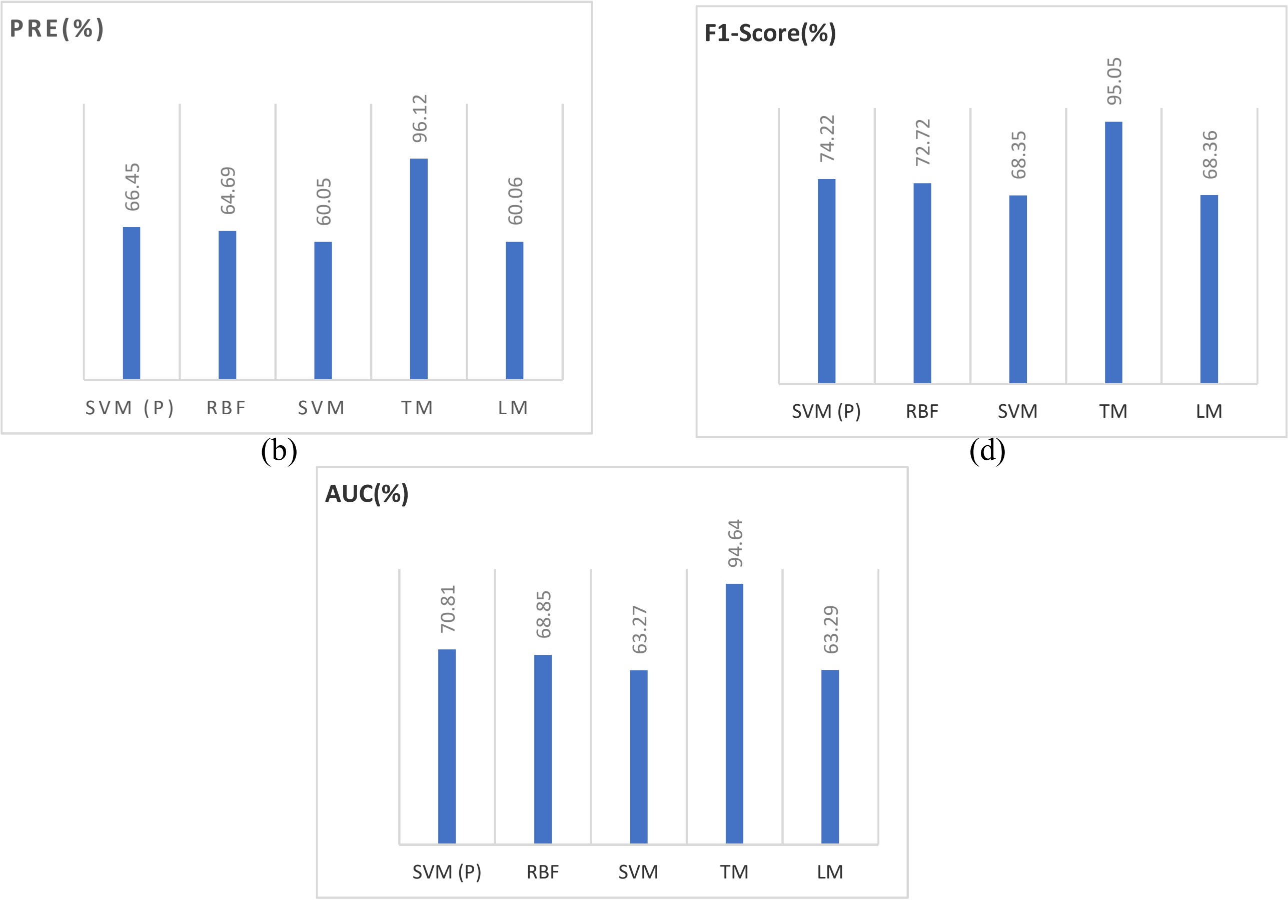
(d)

(e)

FIGURE 2. THE COMPARISON OF SVM(P), RBF, SVM, TM, AND LM USING EPITOPES AND THE DOWNLOADED NEGATIVE DATASET.



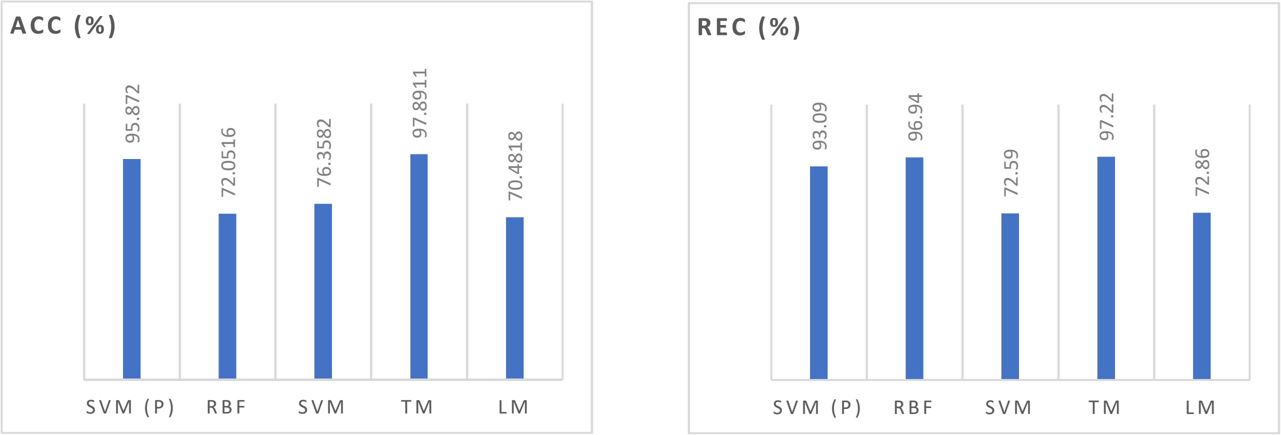
(a) (c)



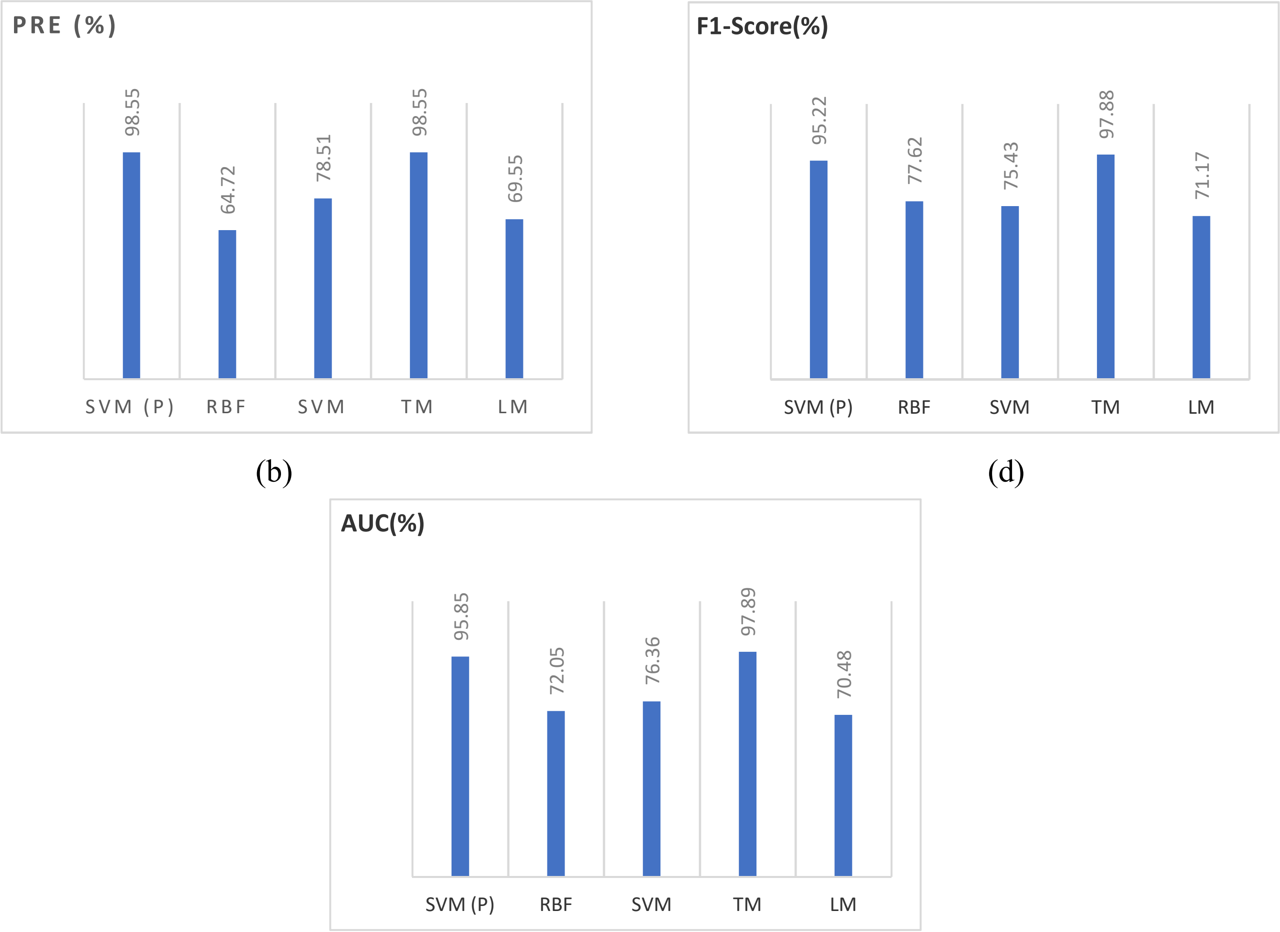
(b) (d)

(e)

FIGURE 3. THE COMPARISON OF SVM(P), RBF, SVM, TM, AND LM USING EPITOPES AND HPRD GENERATED NEGATIVE DATASET



(a) (c)



(b)

(d)

(e)

FIGURE 4. The comparison of SVM(P), RBF, SVM, TM, and LM using COT and HPRD generated negative dataset

3) Experiments using (RandomPairs negative dataset)

In these experiments, we investigate using COT and Epitopes as two features extraction methods while applying the considered machine learning methods on the DIP positive datasets and the generated negative dataset using RandomPairs negative sampling method to study the effect of the number of extracted features on the PPI prediction model on both datasets , and then comparing results of experiments on this negative dataset with results of our generated negative dataset experiments.

Exp5: Applying machine learning methods on the (RandomPairs negative dataset) using COT.

Datasets are converted to COT vectors and concatenated into one matrix (7*98) as an interaction to test the results using 686 features for each interaction using DIP positive dataset and negative dataset generated based on the RandomPairs method. After this experiment was conducted, the following became evident: SVM(P) is the best in all evaluation metrics as 83.8%, 85% ,84.71% for accuracy, recall and AUC respectively, but it also noted that the precision of SVM(P) outperforms SVM, TM, RBF and LM by 4%, 0.5%, 6.5%, 9.7% respectively in varying proportions. F1-score is superior at SVM(P) than SVM, LM, RBF by 4.9%, 7%, 6.9% respectively and with a slight difference from TM by 0.9% as presented in Figure 5. During the experiment, it was noted that SVM, SVM(P) and RBF consumed average half an hour more than TM and LM. 686 features for each interaction were used, making SVM with used kernel functions slow compared to TM and LM. It was also noted that the results of SVM(P) and TM are the best results which proves that features extracted by COT are sufficient for SVM(P) and TM to make the process of classification more accurate than the rest of the classifiers. In general, all methods consumed an average of 30-60 minutes because of the number of features in the huge size of dataset.

Exp6: Applying machine learning methods on the (RandomPairs negative dataset) using Epitopes

Machine learning methods were applied on DIP positive and RandomPairs negative datasets after converting them to Epitopes vectors and concatenated each interaction into one matrix (1*42) to test the results using 42 features for each interaction. As shown in Figure 6, TM has the best accuracy, which outperforms SVM(P), RBF, SVM and LM by 6%, 9.5%, 9.8% and 14.6% respectively. Precision is superior at TM than SVM(P), RBF, SVM and LM by 4.4%, 9.9%, 8.8% and 14.9% respectively. TM achieved the best F1-score, recall and AUC, which were 83.8% ,84.3% and 84% respectively. We also noted that this experiment consumed around 10 minutes Since the size of data is not considered large either, and the number of extracted features was small. We also notice a decrease in accuracy ratios with all used machine learning methods except for TM that achieved accuracy 84%, which demonstrates that extracted features by Epitopes are sufficient for TM to make the classification process more accurate than the rest of the classifiers as shown in Exp2 and Exp3 results. Our investigation in this experiment also reveals that the data nature is overlapping, since the LM approach was the least accurate and this note in all experiments.

4) Experiments using (RecombinePairs negative dataset)

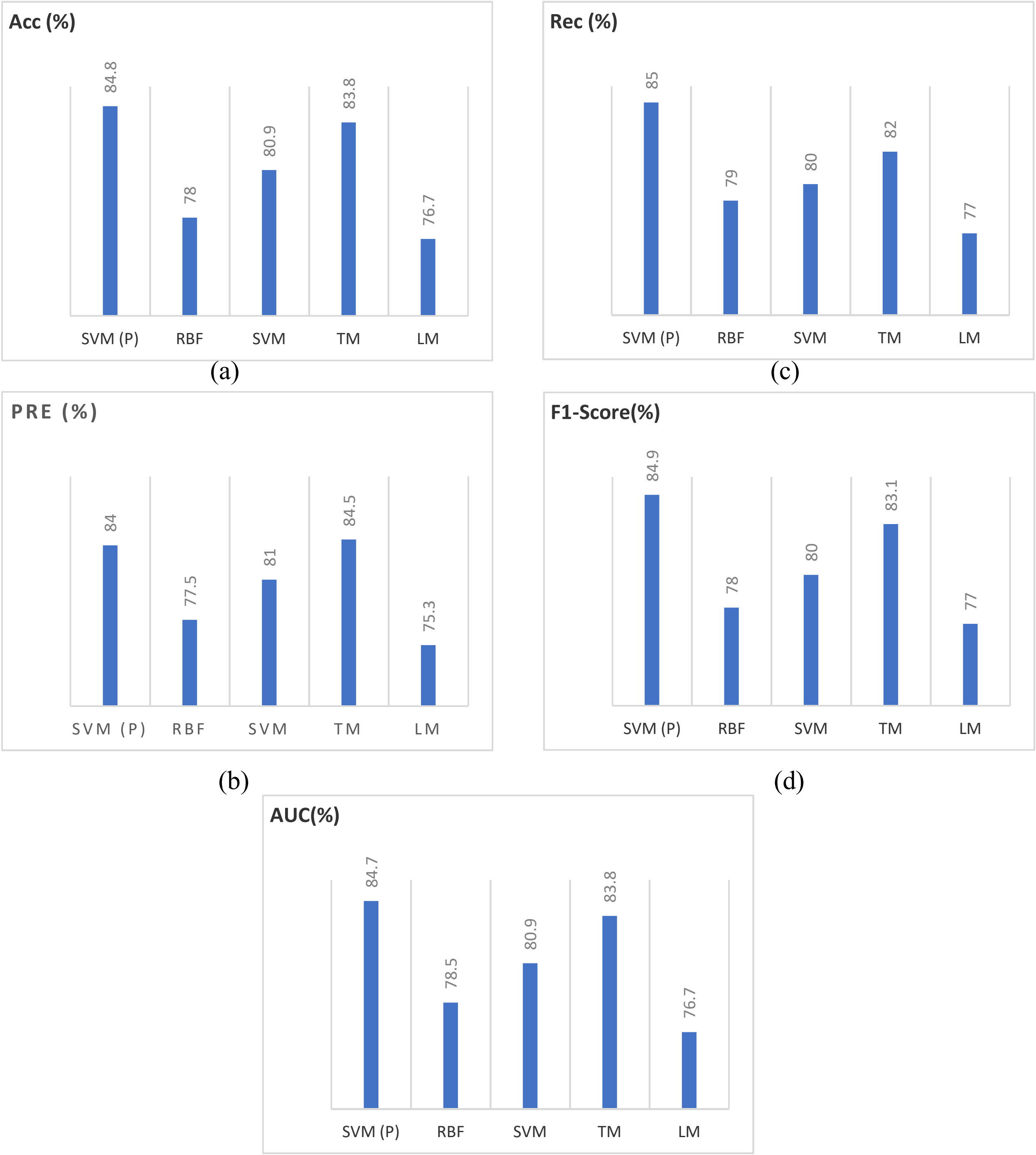
In these experiments, we investigate using COT and Epitopes as two features extraction methods while applying the considered machine learning methods on the DIP positive datasets and the generated negative dataset using RecombinePairs negative sampling method to study the effect of the number of extracted features on the PPI prediction model on both datasets, and then comparing results of experiments on this negative dataset with results of our generated negative dataset experiments.

Exp7: Applying machine learning methods on the (RecombinePairs negative dataset) using COT.

The machine learning methods were applied on DIP positive and RecombinePairs negative datasets after converting them to COT vectors and concatenated into one matrix (7*98) as an interaction to test the results using 686 features for each interaction using the negative dataset generated based RecombinePairs method. As presented in Figure 7, TM and SVM(P) achieved the highest results in comparison to the rest of classifiers which were 81% and 80.2% respectively as accuracy results and in other evaluation metrics results were similar, but TM shows between recall outperforming SVM(P) by 1%, and Recall is superior at TM than RBF, SVM and LM by 9.9%, 5.6% and 16.7% respectively. SVM(P) was better than TM, RBF, SVM, , LM in precision result by 1,7%,10.6%,4.1% and 15.9%. TM outperforms all used classifiers in AUC evaluation metric with results 81.2%,,and SVM(P) achieved best F1-score result which was 80.7% and all evaluation metrics results are represented in Figure 7. This experience did not take long but SVM(P) took about 30 minutes longer than TM, although the results are close in all evaluation metrics. When the negative dataset was changed to RecombinePairs negative datasets. We can also conclude from this experiment that extracted features by COT are enough for TM and SVM(P) to make the classification process, but it becomes clear from the results that extracted features by COT are not enough for the rest of the classifiers for classification process.

Exp8: Applying machine learning methods on the (RecombinePairs negative dataset) using Epitopes.

Machine learning methods were applied on DIP positive and RecombinePairs negative datasets after converting them to Epitopes vectors and concatenated each interaction into one matrix (1*42) to test the results using 42 features for each interaction. In this experiments TM achieved the best accuracy, precision and F1-score results among all used classifiers, which were 80%, 81.3% and 80.3% respectively. AUC is superior at TM than SVM(P), RBF, SVM and LM by 9.8%, 13.8%, 15.3% and 21% respectively. The recall of TM outperforms SVM(P), RBF, SVM and LM by 6.4%, 11.9%, 13.1% and 19.06% respectively, as shown in Figure 8. We also noted that this experiment consumed around 10 minutes Since the size of data is not considered large either and the number of extracted features was small. We also notice that results of TM were significantly higher than the rest of the classifiers which demonstrates that TM can perform the classification process based on extracted features by Epitopes more accurately than the rest of the classifiers.



(a)

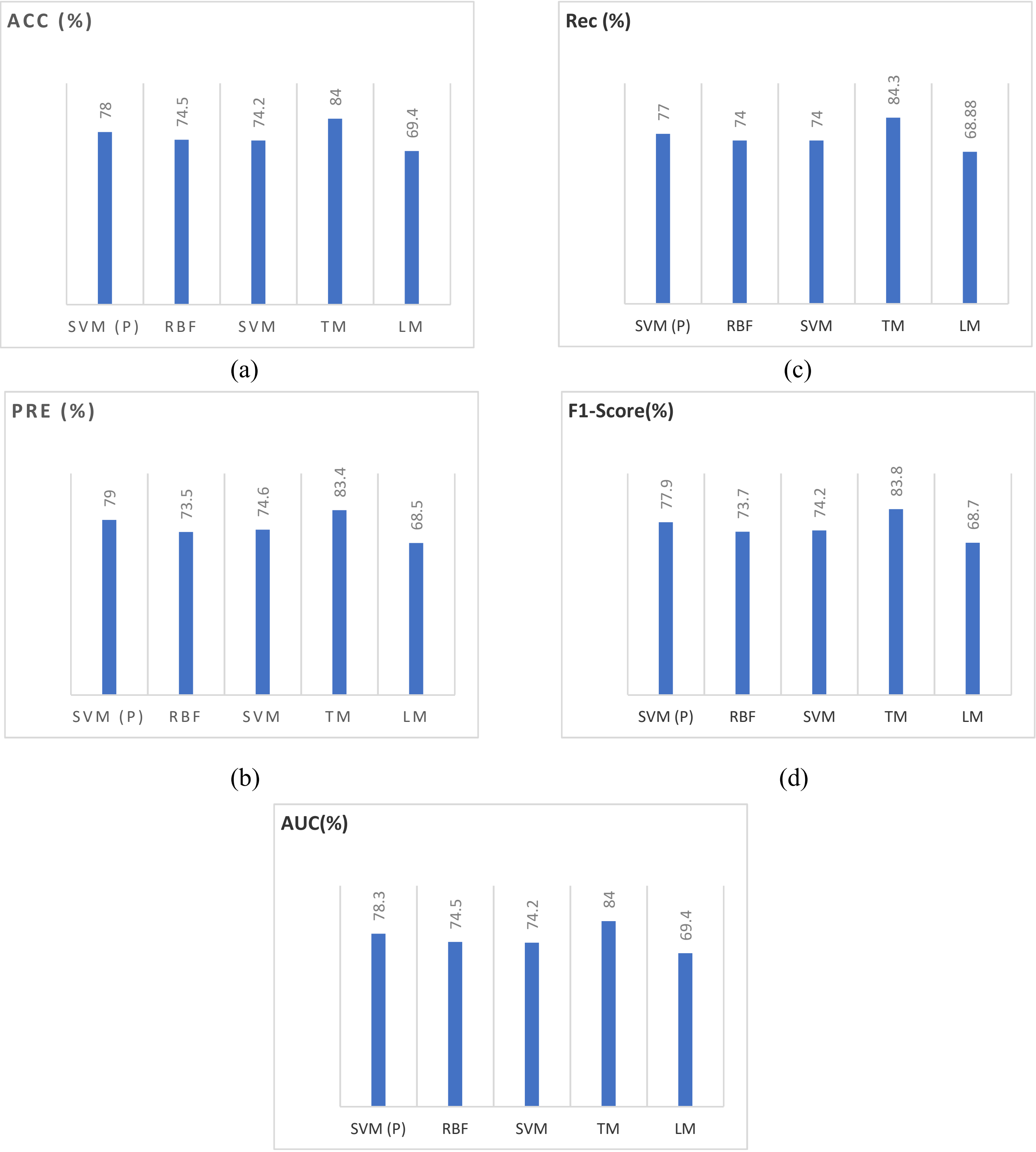
(c)

(b)

(d)

(e)

Figure 5. The Comparison Of Svm(P), Rbf, Svm, Tm, And Lm Using Cot And The Randompairs Negative Dataset



(a)

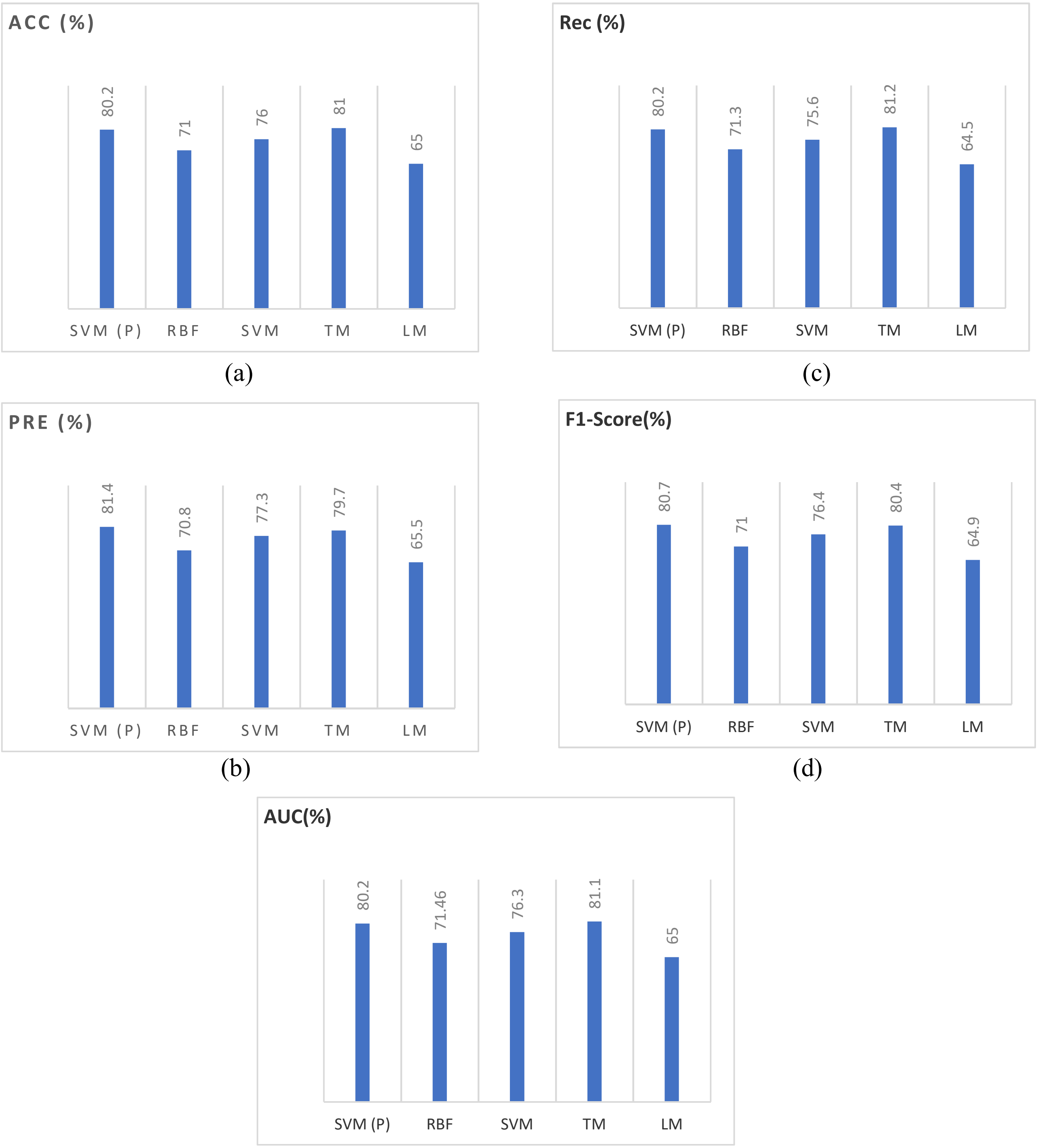
(c)

(b)

(d)

(e)

Figure 6. The Comparison Of Svm(P), Rbf, Svm, Tm, And Lm Using Epitopes And The Randompairs Negative Dataset

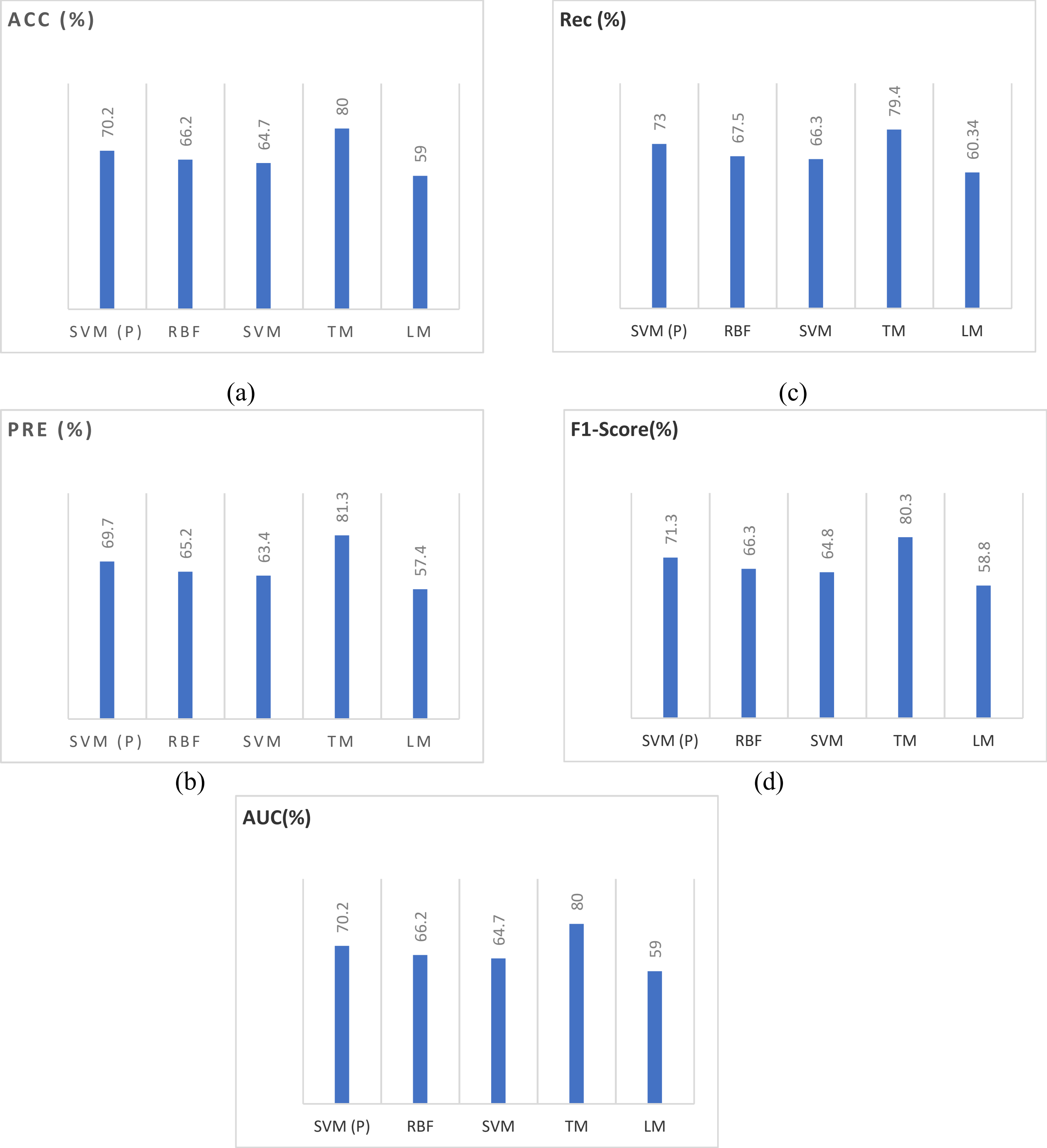


(a) (c)

(b) (d)

(e)

Figure 7. The Comparison Of Svm(P), Rbf, Svm, Tm, And Lm Using Cot And The Recombinepairs Negative Dataset



(a) (c)

(b) (d)

(e)

Figure 8. The Comparison Of Svm(P), Rbf, Svm, Tm, And Lm Using Epitopes And The Recombinepairs Negative Dataset

5) Experiments using (DIP generated negative dataset)

In these experiments, we investigate using COT and Epitopes as two features extraction methods while applying the considered machine learning methods on the DIP positive datasets and the generated negative dataset using the proposed Features-based Negative Generation approach on DIP dataset to study the effect of the number of extracted features on the PPI prediction model on both datasets.

Exp9: Applying machine learning methods on the (DIP generated negative dataset) using Epitopes.

The machine learning methods were applied on the DIP positive and generated negative datasets after converting them to Epitopes vectors and concatenated into one matrix (1*42) as an interaction to test the results using 42 features for each interaction via the negative dataset generated based on the proposed Features-based Negative Generation method. As shown in Figure 9, TM is the best in all evaluation metrics as 86%, 87.2% ,86.02% for accuracy, F1-score and AUC respectively, when the negative dataset was changed to (DIP generated negative dataset) and still using small number of features. Recall is superior at TM than SVM(P), RBF, SVM and LM by 5.5%, 10.3%, 17.5% and 14.5% respectively. The Precision of TM outperforms SVM(P), RBF, SVM and LM by 12%, 13.8%, 16.3% and 13.8% respectively, as shown in Figure 9. It was noted that this experiment took around 8 minutes, because of the small number of extracted features. When the negative dataset was changed to (DIP generated negative dataset) using the same feature extraction method (Epitopes) that was used in Exp6 and Exp8 the accuracy has increased by 2% and 6% comparing with Exp6 and Exp8 respectively, achieving the best accuracy of all experiments using Dip dataset and Epitopes features which was 86% by using TM. It was noted that TM and SVM(P) still the best results by using Epitopes features.

Exp10: Applying machine learning methods on the (DIP generated negative dataset) using COT.

The machine learning methods were applied on the DIP positive and generated negative datasets after converting them to COT vectors and concatenated into one matrix (7*98) as an interaction to test the results using 686 features for each interaction using the negative dataset generated based on the proposed Features-based Negative Generation method. As presented in Figure 10, SVM(P) and TM have the best evaluation metrics results. The differences between them are very simple, but they tend to TM in most of the results. The accuracy of TM outperforms SVM(P), RBF, SVM and LM by 0.5%, 10.5%, 8.5% and 13.5% respectively. SVM(P) consumed an extra average of 4 minute compared to TM. Precision is superior at SVM(P) than RBF, SVM and LM by 12.2%, 9% and 11.9% respectively. Precision of SVM(P) is very slightly superior to TM by 0.4%. TM achieved the best AUC, which was 87.6%. SVM(P) achieved the best F1-score, which was 87.3%. And all experiment results are represented in Figure 10. It was also noted that SVM(P) results are similar to TM in all evaluation metrics and both of them have the best results which proves that COT features fit TM and SVM(P) than the rest of classifiers. When the negative dataset was changed to (DIP generated negative dataset). while using the same feature extraction method (COT) that was used in Exp5 and Exp7 the accuracy has increased by 3% and 7% comparing with Exp5 and Exp7 respectively, achieving the best accuracy of all experiments using Dip dataset and COT features, whereas the accuracy of TM was 87.5%.

5. DISCUSSION ON PERFORMANCE EVALUATION

This section analyzes the obtained results from all experiments and compares the results based on the purpose of each experiment. Firstly, we compare experiments which use the same positive and negative datasets but one of them used Epitopes and the other used COT for feature extraction to demonstrate the effect of using Epitopes. Comparing Exp1 with Exp2 as both of them used the same negative dataset (downloaded negative dataset) and positive dataset (HPRD), V2007 and the result of the comparison showed that the best accuracy in both experiments is converging, and the accuracy of SVM decreased significantly in Exp2 by 7.5% and 22% using SVM and SVM(P) respectively, but accuracy of TM is almost unaffected but reached 92.5%. By comparing Exp3 results with Exp4 results as both of them used negative dataset (HPRD generated negative dataset) and positive dataset (HPRD), V2007, it was found that the difference in accuracy between the best accuracy in both of them was only 3% in favor of COT. However, the accuracy decreased in Exp3 by 13% and 25% using SVM and SVM(P) respectively. But accuracy of TM is almost unaffected and reached 94.5%. Comparing Exp5 results with Exp6 as both of them used negative dataset (RandomPairs generated negative dataset) and positive dataset (DIP). It was found that there is almost no difference in accuracy between the best accuracy in both of them. The accuracy of TM is almost unaffected but reached 84%. the accuracy decreased in Exp6 by 6% using SVM and SVM(P).

By comparing Exp7 results with Exp8 as both of them used negative dataset (RecombinePairs generated negative dataset) and positive dataset (DIP). We found a difference in accuracy of only 1 percent between the best accuracy in both of them. The accuracy of TM and reached 80% and only decreased by 1 percent in favor of COT but decreased significantly by 12% and 10% using SVM and SVM(P) respectively in Exp8. Finally, we compared Exp9 results with Exp10 results as both of them used negative dataset (DIP generated negative dataset) and positive dataset (DIP). We noticed through the result of comparison that the difference in accuracy is 1.5 percent in the case of using TM classifier in favor of COT but decreased by 4.7% and 8% using SVM and SVM(P) respectively in Exp9 and the results of these comparisons were represented in Figure11.

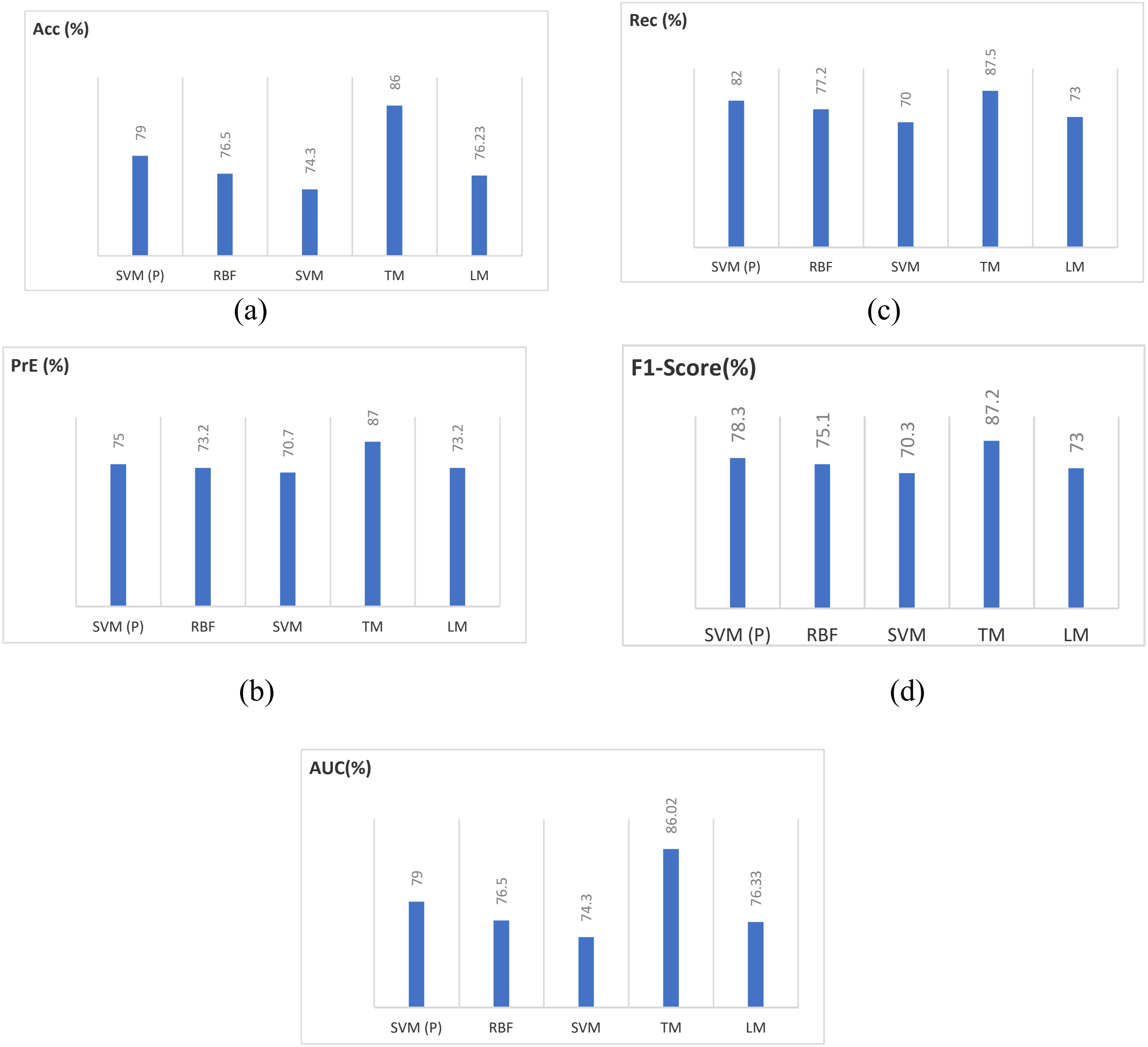
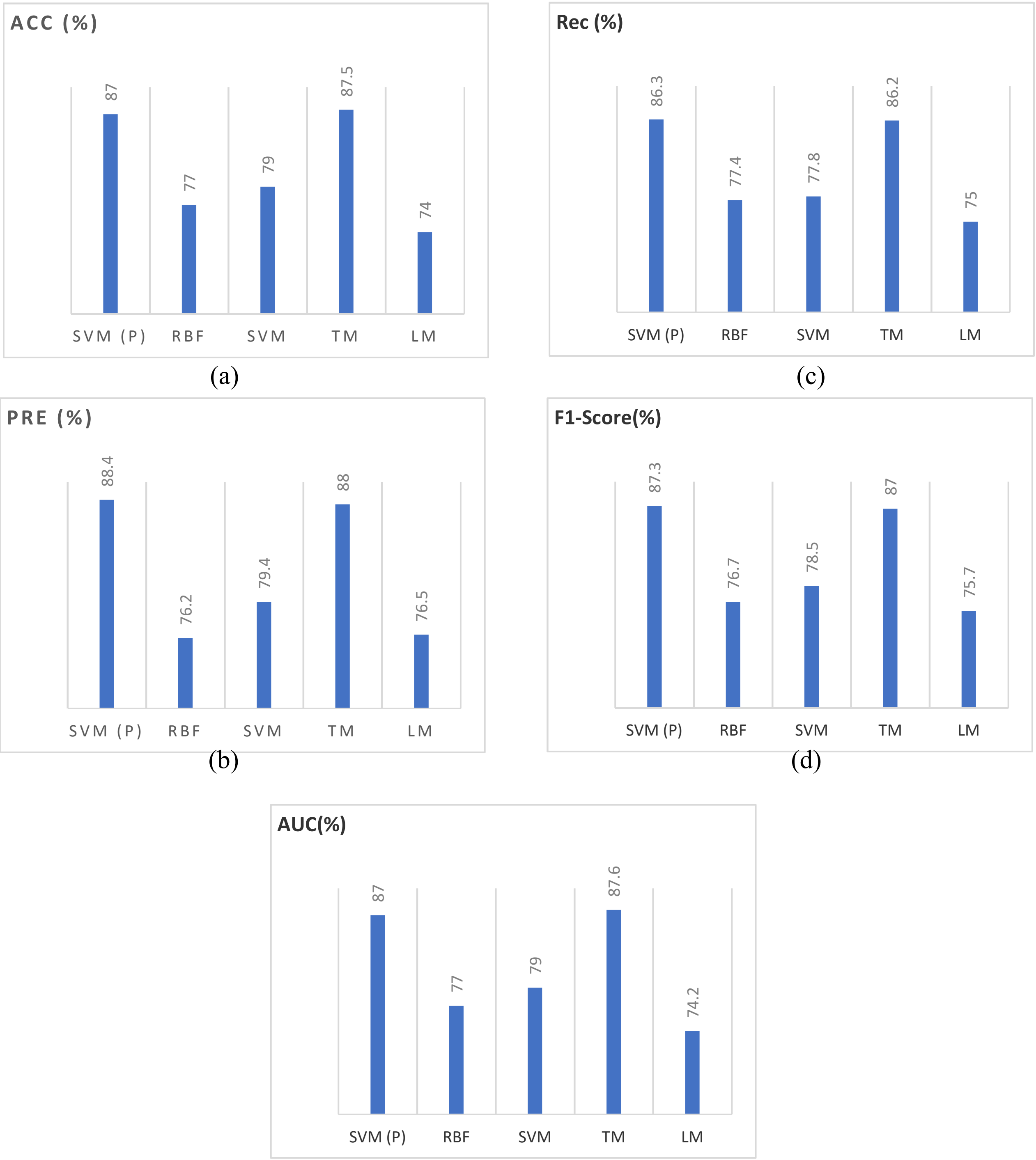


Figure 9. The Comparison Of Svm(P), Rbf, Svm, Tm, And Lm Using Epitopes And Dip Generated Negative Dataset

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(a) (c)

(b) (d)

(e)

Figure 10. The Comparison Of Svm(P), Rbf, Svm, Tm, And Lm Using Cot And Dip Generated Negative Dataset

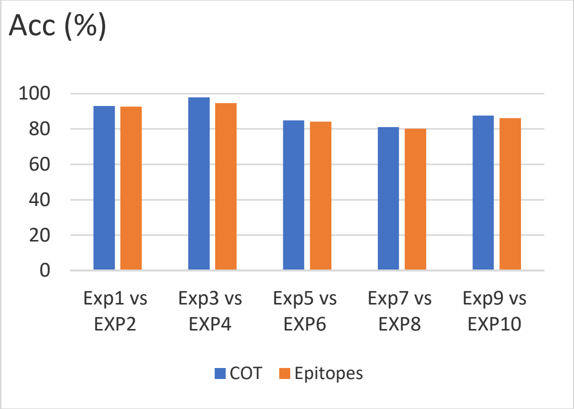


Figure 11. A Comparison Between The Best Accuracy Of Epitopes And The Best Accuracy Of Cot On The Same Datasets.

As a result of what we have observed from the results of comparing those experiments, it was found that the differences between the data used, as well as the feature extraction methods, had an effective role in the achieved results and this is clearly shown in Table 2. The accuracy in all comparisons decreased significantly when using Epitopes for features extraction with all classifiers which indicates that the lower number of features affected the results negatively except for TM, which continued to achieve good results, which indicates that the features extracted by Epitopes are sufficient for classification process with TM classifiers but it is not sufficient for the rest of the used classifiers and this was the aim of comparisons to prove that Epitopes can be used to extract the Features and achieve good accuracy may exceed the accuracy based on COT features. In terms of time, we find the time needed for experiment based on Epitopes features is much less the time needed for experiment based on COT features, where both of them on same dataset and this is due to the large difference in the number of features, for example in Exp1 and Exp2 time of Exp1 is approximately five times the time of Exp2 and all the times needed for the experiments shown in the Table2.

Secondly, we compare experiments to demonstrate the effect of using our proposed Features-based Negative Generation method for negative sampling. When comparing between the results of experiments Exp1 and Exp4, as both of them using COT for features extraction and the same positive dataset (HPRD), V2007 but for negative interactions Exp1 used (downloaded negative dataset) and Exp4 used (HPRD generated negative dataset) which is generated using our proposed Features-based Negative Generation method, we find that the accuracy improved in Exp4 by 4.8%. the same comparing between Exp2 and Exp3, both of them using Epitopes for features extraction and the same positive dataset (HPRD), V2007 but for negative interactions Exp2 used (downloaded negative dataset) and Exp3 used (HPRD generated negative dataset) and we find that the accuracy improved in Exp4 by 2%. We were able to conclude the reason for the superiority of the results based on our PPNIs over the results based on the (downloaded negative dataset), because the (downloaded negative dataset) was selected based on randomness mainly in contrast to our method of extracting PPNIs, and this caused noise in the (downloaded negative dataset), according to research [4], which used the same (downloaded negative dataset). We also tried to compare Features-based Negative Generation with other methods that tried to identify negative dataset, so we compared accuracy of Exp6 and Exp8 with Exp9 where all of them using Epitopes for features extraction and the same positive dataset (DIP) but for negative interactions Exp6 used (RandomPairs negative dataset) , Exp8 used (RecombinePairs negative dataset) and Exp9 used (DIP generated negative dataset) which is generated using our proposed Features-based Negative Generation approach and we find that accuracy has increased in Exp9 by 2% and 6% comparing with Exp6 and Exp8 respectively. By applying the same comparison between Exp5 and Exp7 with Exp10 which also used the same positive dataset (DIP) but for negative interactions Exp5 used (RandomPairs negative dataset), Exp7 used (RecombinePairs negative dataset) and Exp9 used (DIP generated negative dataset) and used COT for features extraction we find that accuracy has increased in Exp10 by 3% and 7% comparing with Exp5 and Exp7 respectively. Here we were also able to conclude the reason for the superiority of the results based on the PPNIs identified by Features-based Negative Generation approach, where both of other two methods RecombinePairs and RandomPairs depend on extracting negative data in the random form as a major step, which causes a noise in the data that affects the accuracy of the distinction between the negative and positive interaction [4,23] and in the event that it is not distinguished if the interaction is negative or positive, both of them considers it negative, which negatively affected the prediction [23].

All of these comparisons proved that accuracy improves when using negative dataset generated by our Features-based Negative Generation method as represented in Table2 and this was the main objective of the research, as this proves that the negative data resulting from our method has less noise than the rest of the negative datasets, which led to a high accuracy of the distinction between negative and the positive data. In terms of time, Epitopes is still less time consuming since when comparing the experiments Exp4 and Exp1, both experiments were the most time consuming within 100%-400% compared to the Exp2 and Exp3 because of the high number of features (686 features) and We mentioned the first four experiments here only because they use the same amount of data and the larger used dataset in our research, so the comparison is accurate. Yet, the best accuracy was achieved using (HPRD dataset) in Exp4 which was 97.8%, where the negative dataset was generated using the proposed Features-based Negative Generation method.

We noticed from our experiments that all experiments which are using COT for features extraction achieved high results with TM and SVM(P) only which means that extracted features by COT are enough for TM and SVM(P) to make the classification process, but based on results extracted features by COT are not enough for RBF, SVM and LM for classification process. And as mentioned before we noticed also most of experiments which are using Epitopes for features extraction achieved highest results with TM which means that extracted features by Epitopes are sufficient for TM to make the classification process more than other classifiers.

Table 2. Summary of the conducted experiments

Experiment	Dataset	Number of Features	Best Classifier and Accuracy	Time consumed (Average in Hours)
Exp1	HPRD positive and downloaded negative datasets	(7*98) matrix	93% using SVM(P) and TM	1-3 hours
Exp2	HPRD positive and downloaded negative datasets	(1*42) vector	92.5% using TM	0.5 hour
Exp3	HPRD Positive and HPRD generated negative datasets	(1*42) vector	94.5% using TM	0.5 hour
Exp4	HPRD Positive and HPRD generated negative datasets	(7*98) matrix	97.8% using TM	0.5-2 hours.
Exp5	DIP Positive and RandomPairs negative dataset	(7*98) matrix	84.8% using SVM(P)	30-60 minutes
Exp6	DIP Positive and RandomPairs negative dataset	(1*42) vector	84% using TM	10 minutes
Exp7	DIP Positive and RecombinePairs negative dataset	(7*98) matrix	81% using TM	30-60 minutes
Exp8	DIP Positive and RecombinePairs negative dataset	(1*42) vector	80% using TM	10 minutes
Exp9	DIP Positive and DIP generated negative datasets	(1*42) vector	86% using TM	8 minutes
Exp10	DIP Positive and DIP generated negative datasets	(7*98) matrix	87.5% using TM	25-50minutes

6. CONCLUSION

In this paper, we tackled the problem of improving Protein-Protein Interactions (PPIs) prediction based on sequence. The “Features-based Negative Generation” method is proposed to generate negative dataset without both random selection that causes noise in selected negative dataset, as well as without alignment that consumes much time for huge datasets. The results of prediction achieved an average accuracy of 97.8% with Conjoint Triad Method (COT) as the feature extraction method. When the same method was used with a randomly generated negative dataset, the accuracy dropped to 93%. Because of the large number of features extracted from COT, which was 343 features, Epitopes was used as another feature extraction method to investigate the effect of the number of features with respect to the prediction accuracy and time consumption. As the number of extracted features was 21 using Epitopes, the accuracy reached 92.5% when applied to a randomly generated negative dataset. However, using our generated negative dataset, the accuracy has increased to 94.5%. Besides, the time consumed was 80-83% less than the same experiment using COT. Thus, the PPIs prediction is more accurate with an average of 2-4% using the generated negative dataset, while considering Epitopes for feature extraction, which was proved to be is faster than COT by an average of 80-83%, due to the fewer number of features

7. FUTURE WORK

We intend to investigate novel properties of proteins, such as structural and ontology features, that could improve prediction accuracy. We also will applay new negative generation method depending on protein structural information and use extracted features by Epitopes in drug target interaction prediction and use the predicted PPIs in the preddiction of protein function depending on similarity in iteractions. Also we will use PPIs in protein diseases prediction.

REFERENCES:

[1] Zahiri J, Bozorgmehr J, Masoudi-Nejad A. Computational prediction of protein–protein interaction networks: Algorithms and resources. Current Genomics. 2013;14(6):397–414.

[2] Skrabanek L, Saini HK, Bader GD, Enright AJ. Computational prediction of protein–protein interactions. Molecular Biotechnology. 2007;38(1):1–17.

[3] Tripathi LP, Chen Y-A, Mizuguchi K, Murakami Y. Network-based analysis for Biological Discovery. Encyclopedia of Bioinformatics and Computational Biology. 2019;:283–91.

[4] Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, et al. Predicting protein–protein interactions based only on sequences information. Proceedings of the National Academy of Sciences. 2007;104(11):4337–41.

[5] Francisquini R, Berton R, Soares SG, Pessotti DS, Camacho MF, Andrade-Silva D, et al. Community-based network analyses reveal emerging connectivity patterns of protein-protein interactions in murine melanoma secretome. Journal of Proteomics. 2021;232:104063.

[6] Barh D, Yiannakopoulou EC, Salawu EO, Bhattacharjee A, Chowbina S, Nalluri JJ, et al. In silico disease model: From simple networks to complex diseases. Animal Biotechnology. 2020;:441–60.

[7] Eid F-E, ElHefnawi M, Heath LS. Denovo: Virus-host sequence-based protein–protein interaction prediction. Bioinformatics. 2015;32(8):1144–50.

[8] Hao T, Wang Q, Zhao L, Wu D, Wang E, Sun J. Analyzing of molecular networks for Human Diseases and Drug Discovery. Current Topics in Medicinal Chemistry. 2018;18(12):1007–14.

[9] Fields S, Song O-kyu. A novel genetic system to detect protein–protein interactions. Nature. 1989;340(6230):245–6.

[10] Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, et al. Functional Organization of the yeast proteome by systematic analysis of protein complexes. Nature. 2002;415(6868):141–7.

[11] Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams S-L, et al. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature. 2002;415(6868):180–3.

[12] Göktepe YE, Kodaz H. Prediction of protein-protein interactions using an effective sequence based combined method. Neurocomputing. 2018;303:68–74.

[13] Sarkar D, Saha S. Machine-learning techniques for the prediction of protein–protein interactions. Journal of Biosciences. 2019;44(4).

[14] Lian X, Yang S, Li H, Fu C, Zhang Z. Machine-learning-based predictor of human–bacteria protein–protein interactions by incorporating comprehensive host-network properties. Journal of Proteome Research. 2019;18(5):2195–205.

[15] Romero-Molina S, Ruiz-Blanco YB, Harms M, Münch J, Sanchez-Garcia E. PPI-detect: A support vector machine model for sequence-based prediction of protein-protein interactions. Journal of Computational Chemistry. 2019;40(11):1233–42.

[16] Yang X, Yang S, Li Q, Wuchty S, Zhang Z. Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Computational and Structural Biotechnology Journal. 2020;18:153–61.

[17] Zhang J, Zhu M, Qian Y. Protein2vec: Predicting protein-protein interactions based on LSTM. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2020;:1–.

[18] Sumonja N, Gemovic B, Veljkovic N, Perovic V. Automated feature engineering improves prediction of protein–protein interactions. Amino Acids. 2019;51(8):1187–200.

[19] Baichoo S, Ouzounis CA. Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. Biosystems. 2017;156-157:72–85.

[20] Sedgewick R, Flajolet A. Introduction to the design & analysis of algorithms. Boston: Pearson Addison-Wesley; 2011.

[21] Bairoch A, Apweiler R. The Swiss-PROT Protein Sequence Data Bank and its supplement TREMBL. Nucleic Acids Research. 1997;25(1):31–6.

[22] Zhang L, Yu G, Guo M, Wang J. Predicting protein-protein interactions using high-quality non-interacting pairs. BMC Bioinformatics. 2018;19(S19).

[23] Ben-Hur A, Noble WS. Choosing negative examples for the prediction of protein-protein interactions. BMC Bioinformatics. 2006;7(S1).

[24] Wei L, Xing P, Zeng J, Chen JX, Su R, Guo F. Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. Artificial Intelligence in Medicine. 2017;83:67–74.

[25] Chen C, Zhang Q, Ma Q, Yu B. LIGHTGBM-PPI: Predicting protein-protein interactions through LIGHTGBM with multi-information fusion. Chemometrics and Intelligent Laboratory Systems. 2019;191:54–64.

[26] Fang H, Zhong C, Tang C. Predicting protein-protein interactions between banana and fusarium oxysporum race 4 integrating sequence and domain homologous alignment and neural network verification. 2021;

[27] Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. Hippie v2.0: Enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Research. 2016;45(D1).

[28] Bader GD. Bind--the Biomolecular Interaction Network Database. Nucleic Acids Research. 2001;29(1):242–5.

[29] Mewes HW. MIPS: A database for genomes and protein sequences. Nucleic Acids Research. 2000;28(1):37–40.

[30] Xenarios I. Dip, the database of interacting proteins: A research tool for studying cellular networks of protein interactions. Nucleic Acids Research. 2002;30(1):303–5.

[31] Jiankuan Ye, Kulikowski C, Muchnik I. Sequence-based protein-protein interaction prediction optimized for target selection in biological experiments. 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. 2005;

[32] Zhou H, Gao S, Nguyen NN, Fan M, Jin J, Liu B, et al. Stringent homology-based prediction of H. Sapiens-M. tuberculosis H37RV protein-protein interactions. Biology Direct. 2014;9(1).

[33] Ben-Hur A, Weston J. A user's guide to support Vector Machines. Methods in Molecular Biology. 2009;:223–39.

[34] Zhou H, Rezaei J, Hugo W, Gao S, Jin J, Fan M, et al. Stringent DDI-based prediction of H. Sapiens-M. tuberculosis H37RV protein-protein interactions. BMC Systems Biology. 2013;7(S6).

[35] Evans P, Dampier W, Ungar L, Tozeren A. Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs. BMC Medical Genomics. 2009;2(1).

[36] Northey TC, Barešić A, Martin AC. IntPred: A structure-based predictor of protein–protein interaction sites. Bioinformatics. 2017;34(2):223–9.

[37] Wang J, Zhang L, Jia L, Ren Y, Yu G. Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences. International Journal of Molecular Sciences. 2017;18(11):2373.

[38] Barkat MR, Moussa SM, Badr NL. Drug-target interaction prediction using machine learning. 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS). 2021;

[39] Ghosh, B. and Parker, A. Project final: epitope classification using support vector machines. 2009.

[40] Liu S, Liu C, Deng L. Machine learning approaches for protein–protein interaction hot spot prediction: Progress and comparative assessment. Molecules. 2018;23(10):2535.

[41] Seven most popular SVM kernels [Internet]. Dataaspirant. 2020 [cited 2021Dec12]. Available from: https://dataaspirant.com/svm-kernels

[42] Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An introduction to decision tree modeling. Journal of Chemometrics. 2004;18(6):275–85.

[43] Patel N, Singh D. An algorithm to construct decision tree for machine learning based on similarity factor. International Journal of Computer Applications. 2015;111(10):22–6.

[44] Park Y, Marcotte EM. Revisiting the negative example sampling problem for predicting protein-protein interactions. Bioinformatics. 2011;27(21):3024–8.

[45] M. Sayed Barkat, Sherin Moussa, "Protein-Protein Negative Interaction Dataset", IEEE Dataport, 2022, doi: https://dx.doi.org/10.21227/51xw-5q83.

[46] UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics. Uniprot Consortium [Internet]. UniProt ConsortiumEuropean Bioinformatics InstituteProtein Information ResourceSIB Swiss Institute of Bioinformatics. [cited 2022Jan18]. Available from: https://www.uniprot.org/

[47] Smialowski P, Doose G, Torkler P, Kaufmann S, Frishman D. Proso II - A new method for protein solubility prediction. FEBS Journal. 2012;279(12):2192–200.

[48] Liu L, Zhu X, Ma Y, Piao H, Yang Y, Hao X, et al. Combining sequence and network information to enhance protein–protein interaction prediction. BMC Bioinformatics. 2020;21(S16).

[49] Zhou G, Wang J, Zhang X, Guo M, Yu G. Predicting functions of maize proteins using graph convolutional network. BMC Bioinformatics. 2020;21(S16).

[50] Zhang S, Duan X. Prediction of protein subcellular localization with oversampling approach and Chou's general pseaac. Journal of Theoretical Biology. 2018;437:239–50.

[51] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition. 1997;30(7):1145–59.

[52] Manavalan B, Shin TH, Lee G. PVP-SVM: Sequence-based prediction of phage virion proteins using a support vector machine. Frontiers in Microbiology. 2018;9.