IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 6, 2020, accepted January 5, 2021, date of publication January 12, 2021, date of current version January 22, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3051006

Determining Protein–Protein Interaction Using Support Vector Machine: A Review

ARIJIT CHAKRABORTY[1], SAJAL MITRA[2], DEBASHIS DE[3], (Senior Member, IEEE), ANINDYA JYOTI PAL[4], FERIAL GHAEMI[5], ALI AHMADIAN[6,7], (Member, IEEE), AND MASSIMILIANO FERRARA[7]

[1]Department of Computer Application, The Heritage Academy, Kolkata 700107, India
[2]Department of Computer Science and Engineering, Heritage Institute of Technology, Kolkata 700107, India
[3]Department of Computer Science and Engineering, West Bengal University of Technology, Kolkata 700064, India
[4]The University of Burdwan, Bardhaman 713104, India
[5]Department of Chemical and Process Engineering, Faculty of Engineering and Built Environment, The National University of Malaysia (UKM), Bangi 43600, Malaysia
[6]Institute of Industry Revolution 4.0, The National University of Malaysia (UKM), Bangi 43600, Malaysia
[7]DECISIONS Lab, University Mediterranea of Reggio Calabria, 89124 Reggio Calabria, Italy

Corresponding authors: Arijit Chakraborty (arijit.chakraborty@heritageit.edu) and Ali Ahmadian (ahmadian.hosseini@gmail.com)

This work was supported in part by the DECISIONS Lab, Mediterranea University of Reggio Calabria, Italy.

ABSTRACT Protein-Protein Interaction (PPI) is a network of protein interconnections which regulates most of the biological methods. A sound state of biota largely depends on synchronized interactions between protein molecules, and any aberrant interactions between protein molecules may lead to complications, including cervical leukemia, tuberculosis, and other neural disorders. In PPI investigation, a plethora of computational methods have been developed over the years to analyze and predict PPI conclusively; however, a majority of these techniques proved to be strenuous and expensive. Therefore, the need for faster, accurate, and critical analysis of PPI warrants the adoption of Machine Learning (ML) methods such as Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest Model (RFM). These classifiers are useful in PPI unfolding in terms of amino acid sequence data. The SVM classifier, in particular, is serviceable in solving a majority of complex classification problems producing robust results in a reasonable time frame. This publication summarizes some state-of-art SVM based PPI investigations and challenges incurred in the application of the SVM method.

INDEX TERMS Artificial neural network, machine learning, protein-protein interaction, support vector machine.

I. INTRODUCTION

Proteins are macromolecules consisting of long strings of amino acid residues that perform several functions inside organisms, including replication of DNA, stimuli-response mechanism, and molecular transportation. Biological processes follow a concerted mechanism in which several protein molecules participate where DNA molecules sustain necessary biological information, which expresses through functions of proteins molecules. PPIs are the bio-physical connections of high specificity set between protein molecules produced by biochemical phenomenon. The PPI constitutes cellular communications in living beings that take place through the exchange of signals, which may come either from an environment or from neighboring cells. These signals bind to receptor proteins to reach the desired cell through a channel known as cell membranes. These receptors are connected inside and outside of a cell, forming a signaling pathway between source and destination. Communication between proteins such as tissue proteins, proteins from viruses, microorganisms, and bacteria resulted in disease-causing mutations. Therefore, it is imperative to analyze protein communications that help in the identification of such mutations. These mutations either affect binding interfaces or causes biochemical impairment by amending the job of an enzyme. The PPI analysis empowers us with the following:

- Detection of protein complexes.
- Identification of domain interactions.

The associate editor coordinating the review of this manuscript and approving it for publication was Xinyu Du.

- The involvement of protein in disease pathways.
- Developing effective strategies in drug design.

Conventional bio-physical methods for PPI identification are both tedious and expensive. In contrast, traditional computational methods are confined by pre-requisite knowledge about gene-neighbourhoods, phylogenetic sketches, and sequence interpretation to render favourable PPI prediction outcomes. Machine learning (ML) methods, including SVM, ANN, RFM, and deep-learning, deliver critical means for judicious prediction of PPIs based on the direct derivation of protein information from amino acid sequences. In this context, Xia *et al.* [1] reviewed the adoption of computational methods in Genomic, structure, domain, and sequence-based approaches. The SVM method offers multi-faceted aspects, such as integrating statistical descriptors with binary coding of protein sequences, and operative use of SVM variant, i.e., two-class SVM on heterogeneous protein complexes, both stable and temporary, in PPI recognition. Reference [1] reviewed the use of the Rotation Forest on sequence-based approach, wherein the SVM method is equally useful.

These constituents influenced us to review PPI prediction through the lenses of SVM. Consequently, we reviewed SVM's performance based on the cluster, genome, domain and customized feature-encoding tool.

II. PRELIMINARIES

The study of PPI can be conceptualized from diverse perspectives. In computer science, a PPI system is modelled as a graph $G = (V, E)$, where V is the set of the protein vertices, and E is the set of the edges representing pairwise protein interactions. Weighted edges in graph G used to describe reliability information associated with such interactions. A PPI system can also be considered as a network of interconnected nodes, which build a global network of protein interaction architecture. The PPI network system is useful in depicting, visualizing, and quantifying cellular functions. In this context, the authors of [2] used Graph Fragmentation Algorithm (GFA) derived and adapted from the Max Flow Algorithm (MFA) to identify protein complexes in the PPI network. In their work, authors of [3] used a graph mining algorithm for determining protein interactions by merging local cliques to obtain maximal dense regions. In contrast, authors of [4], proposed a cost clustering algorithm for predicting protein complexes, in which the entire PPI network partitioned into clusters for searching interacting neighbours. Authors of [5], suggested the application of spectral graph method in unhiding the topological structures consisting of similar functional groups.

The biophysical methods encompass a plethora of techniques to describe, recognize, and predict PPI. These methods are useful in analyzing the bio-molecular roles of PPI at the atomic level. The organization of biophysical methods is shown in Fig. 1.

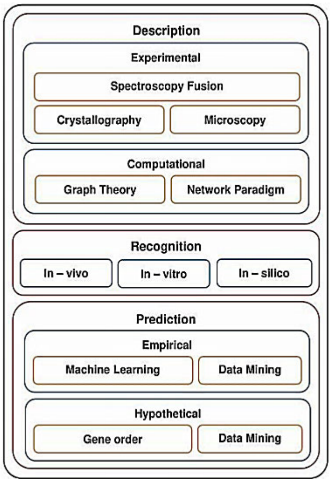


FIGURE 1. PPI identification methods.

Rao *et al.* [6] reviewed the role of biological methods for detecting PPI. These methods are primarily classified into three basic types' in-vitro [7], [8], in-vivo [9], and in-silico [10]. The TAP tagging method is an in-vitro approach to investigate yeast interactome. Tap tagging a three-stage method, where a two-fold tagging of protein to its chromosomal locus done, followed by double purification using SDS PAGE and the proteins that remain affiliated to target proteins are scrutinized. Finally, mass spectrometry analysis is used to detect PPI. Another method, namely, Affinity Chromatography, is used to detect weak protein interactions at the molecular level. However, it has a drawback of generating false-positive results due to a high particularity between protein molecules. To overcome the limitation of Affinity Chromatography, a hybrid approach using affinity chromatography with SDS-PAGE and mass spectrometry employed for detecting PPI reasonably, whereas, the Protein-fragment Complementation Assay (PCA) method is useful for identifying interaction among proteins with varying molecular masses. A comparison between in-vitro, in-vivo, and in-silico methods is listed in Table 1.

TABLE 1. PPI recognition methods.

Method	Area	Application	Limitation
In-vitro [7-8]	Experiments conducted in controlled conditions that are external to living organisms.	Caco-2 cell tests. It is used to measure the absorption of compounds of the gastrointestinal instances. [7]	In-vitro techniques may not distinguish 99% of varieties in the human micro biota. [8]
In-vivo [9]	Experiments on living organisms.	Discovery of the formulations of explicit drugs set and their behaviors. [9]	A shortfall of offering immediate benefit with long term impairment.
In-silico [10]	Performed in a simulated environment.	Used in Sequencing, Molecular Modeling, and Whole-cell facsimile. [10].	Simulated molecular dynamics and simplified assumptions.

The binding affinity between protein molecules largely depends on the presence of a small fraction of the residues in the protein-protein interfaces [11]–[13]. These critical residues are generally regarded as 'Hotspots Positions', which specify mutational spots where the increase in free binding energy >= 2.0 kcal/mol. A method, namely, Alanine Scanning Mutagenesis (ASM), is widely used for Hotspots detection to identify primary intonation in PPI [14], [15]. An extensive analysis of Hotspots-PPI helps in unfolding leukemic genes [16]. In this context, some Hotspot based PPI study along with results listed in Table 2.

TABLE 2. Hot spot prediction result.

Authors	Database	Dataset	Results
Lise S. et al. [17]	Alanine Scanning Energetics database (ASEdb) [15], and Protein Data Bank (PDB) [18]	81 Hot Spots, and 268 non-Hot Spots	Precision 56% Recall 65%
Tuncbag N. et al. [19]	Binding Interface Database (BID) [20]	54 Hot Spots, and 58 non-Hot Spots	Accuracy 70%, Precision 73% and Recall 59%
Lise S. et al. [21]	[15,18]	81 Hot Spots, and 268 non-Hot Spots	Precision 61% Recall 69%
Qiao Y. et al. [22]	[15,20]	62 Hot Spots, and 92 non-Hot Spots	F-measure 62% Recall 82%

The biophysical methods are often time-consuming and labour-intensive, and these large-scale experiments usually suffer from high false-positive rates [23]. Consequently, the ML classification algorithms are extended in the PPI study to realize and predict protein interactions effectively.

A. MACHINE LEARNING

Machine learning is a part of Artificial Intelligence (AI) that assists machines to spontaneously comprehend and learn from experience about a given dataset for making accurate predictions without the need to be explicitly programmed. The description of two machine learning methods Support Vector Machine, Artificial Neural Network, and Confusion Matrix used to analyse the classification performance problems are as follows:

1) SUPPORT VECTOR MACHINE

In their work, Vapnik *et al.* [24] introduced the theory of SVM classification. In [24], a hyper-plane is created to classify voluminous data in appropriate classes in a reasonable time frame shown in Fig. 2. In Fig. 2. the dotted line represents the hyper-plane, which separates the data points into two classes.

The data points that are nearest to the hyper plane are called support vectors. The quality of SVM classification depends on maximizing the margin between class data points from the hyper-plane. The proportion of unambiguously identified protein interaction data is minute compared to the data of diverse organisms, wherein, the application of SVM can be elongated to classify PPI prudently. Description of mathematical background of SVM is given below.

In SVM, x_i is an element in the input space X, i.e., $x_i\epsilon$ X, and y_i is an element in the output space Y, i.e., $y_i\epsilon$ Y{−1, +1}, where −1 and +1 represent two different class labels of output space Y. If y_i is the corresponding class of x_i, then the pair {x_i, y_i} is used to train the SVM method. For a given weight vector w, the linear separation of the input data is done by exercising (1) and (2), where. represents the vector multiplication.

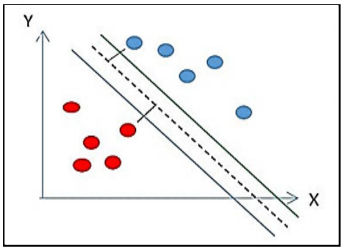


FIGURE 2. SVM classification.

$$w^T .X + B > 0\, for\ y_i = +1 \tag{1}$$
$$w^T .X + B \le 0\, for\ y_i = -1 \tag{2}$$

In (1) and (2), B is the bias in the SVM method. The matching decision function is given in (3).

$$w^T .X + B = 0 \tag{3}$$

2) ARTIFICIAL NEURAL NETWORK

Artificial Neural Network [25] method is also faster and accurate in identifying protein interactions. This model simulates the human brain and consists of three primary layers, an input layer, hidden layers, and the output layer shown in Fig. 3.

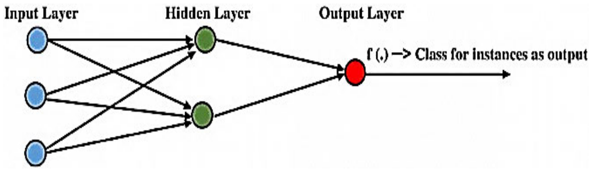


FIGURE 3. ANN classification.

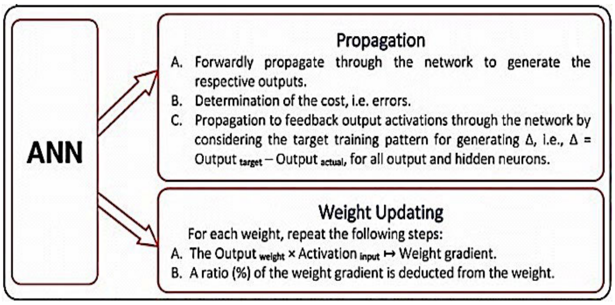
In ANN, each neuro-signal is interconnected by weighted edges, and the activation function provides an output value corresponding to incoming signals. The concept of the Back-propagation algorithm used in Multi-Layer Perceptron (MLP), i.e., ANN is a two-step process shown in Fig. 4.

In [24], [25], the role of kernel function is critical. A kernel function uses a linear classifier to unravel a nonlinear problem. It involves mapping of linearly non-separable instances into a higher N-dimensional plane to make them linearly separable. For input vectors $\overrightarrow{y_1}\,\overrightarrow{y_2}$, kernel functions listed in Table 3.

3) CONFUSION MATRIX

A confusion matrix is the summary of the quality of the solution for a given classification problem. The confusion matrix demonstrates how much a classification model is confused while making predictions. Some commonly used terms in the confusion matrix are listed in Table 4.

The performance of a classification model depends on the proportion of data it can correctly classify. In Table 5, classification performance measures are listed.



A. Forwardly propagate through the network to generate the respective outputs.
B. Determination of the cost, i.e. errors.
C. Propagation to feedback output activations through the network by considering the target training pattern for generating Δ, i.e., Δ = Output$_{target}$ − Output$_{actual}$, for all output and hidden neurons.

For each weight, repeat the following steps:
A. The Output$_{weight}$ × Activation$_{input}$ ↦ Weight gradient.
B. A ratio (%) of the weight gradient is deducted from the weight.

FIGURE 4. Two-step process in ANN.

TABLE 3. Kernel functions.

Kernel	Description	Parameter	Suitability
Linear	$K(x,y) = x^T y + C$	C is the constant	Easy to implement with faster response.
Poly Kernel	$K_p(\overrightarrow{y_1},\overrightarrow{y_2}) = (\overrightarrow{y_1}^T,\overrightarrow{y_2} + 1)^d$	d is the degree of polynomial	The relationships among PPI datasets depend on the order of the polynomial.
Normalized Poly kernel	$K_{normp}(\overrightarrow{y_1},\overrightarrow{y_2}) = \frac{K_p(\overrightarrow{y_1},\overrightarrow{y_2})}{\sqrt{(\overrightarrow{y_1}.\overrightarrow{y_1})}\sqrt{(\overrightarrow{y_2}.\overrightarrow{y_2})}}$		
Radial Basis Function (RBF)	$K_{RBF}(\overrightarrow{y_1},\overrightarrow{y_2}) = e^{-\gamma\lVert\overrightarrow{y_1}-\overrightarrow{y_2}\rVert^2}$	ϒ is the exponent parameter	The data classification based on circles or hyper-spheres.

TABLE 4. Confusion matrix.

Term	Observation	Prediction
True Positive (TP)	Positive	Positive
False Positive (FP)	Negative	Positive
False Negative (FN)	Positive	Negative
True Negative (TN)	Negative	Negative

TABLE 5. Performance measures of ML based classifiers.

Term	Description	Measure
Accuracy	The proportion of correctly classified data instances.	$\frac{TP + TN}{TP + FP + FN + TN}$
Recall	The proportion of correctly classified positives to total no. of positives.	$\frac{TP}{TP + FN}$
Precision	The proportion of correctly classified positives to total no. of positive predictions.	$\frac{TP}{TP + FP}$
F-Measure	A Harmonic Mean of Recall and Precision.	$\frac{2 \times Precision \times Recall}{Precision + Recall}$

B. PPI DATABASE

The recent development in bioinformatics braced thescientific community in the acquisition and storing of PPI data in databases. These databases contain gene information, i.e., cellular functions, structures, sequences, and species-specific information in a computer-readable form.

Lac operon of E. coli dataset [26] stores the genetic information of lactose metabolism, which terminates as a repressor in the presence of lactose. An operon is a part of DNA in which genes, i.e., structural or regulatory or operator genes, are placed adjacently, and it constitutes the functional unit that provides transcription and regulatory activities. The Database Negatome [27] and General Repository for Interaction Datasets (GRID) [28] used to determine Interacting Proteins (IP) and Non-Interacting Proteins (NIP) with features. [27] stores NIP, whereas [28] is a web-based platform that caters both IPs and NIPs of Saccharomyces Cerevisiae. However, a significant drawback of [27], [28] is that both of them do not provide attributes, and to overcome this limitation, KUPS (The University of Kansas Proteomics Service) database [29] developed. It stores ready to use high throughput IP and NIP data for ML of Kansas Proteomics Service) database [29] developed. It stores ready to use high throughput IP and NIP data for ML methods.

The Munich Information Centre for Protein Sequences (MIPS) [30] is a secondary database of manually curated PPI data of mammals and high-quality genome data of Saccharomyces Cerevisiae, NeurosporaCrassa, and Arabidopsis Thaliana. It is a public-database catering to the need of a wide variety of users through a user-friendly interface supported by easily executable query languages for retrieving pieces of information, thereby providing an effective search mechanism for finding protein data of interest. The Database of Interacting Proteins (DIP) [31] stores PPI data generated from both manually curated sources and automatically via computational methods, and the vast PPI information in [31] is used to create reliable PPI set in a single form. Kyoto Encyclopaedia of Genes and Genomes (KEGG) [32] is an assembly of databases of genomes, diseases, and drug ingredients. This database is a huge source of information to identify the high-level functions and class values of different organisms. It provides bio-molecular information and large-scale molecular information created by genome arrangement using high-throughput investigational skills. The GenBank database [33] is a collection of organism's genetic sequences and translations of protein sequences. It is an open-access database and is a part of the International Nucleotide Sequence Database Collaboration. Some widely recognized PPI databases are listed in Table 6.

III. PPI CLASSIFICATION

This publication mainly incorporates a review of some of the state-of-the-art PPI classifications using the SVM variants. However, both SVM and ANN methods are useful in heterogeneous classification problems. The SVM's performance depends on kernel choice, whereas the neural net's performance relies on activation function. Therefore, these methods are not offbeat in the tasks they perform except in approaches and implementations. Consequently, this manuscript reviewed a PPI study with SVM and ANN to realize and illustrate SVM and ANN's performance in recognizing interacting and non-interacting amino acid pairs

TABLE 6. PPI databases.

Database	Version	Type	Data-type	Organism	Unique Features	Dataset Information
Negtome [27]	2.0	Secondary	Experimental & Predicted	Multiple	Information on non-interacting protein pairs, physical annotation, and protein structure	2171 Mammalian proteins, 4397 Proteins from PDB database, 1234 Proteins from PDB & PFAM, 6532 Mammal & PDB proteins
KUPS [29]	-	Secondary	Experimental	Multiple	Data collected from IntAct, HPRD, MINT, UniProt, and Gene Ontology databases	185446 IPs, and 1.5 billion NIPs
MIPS[30]	-	Secondary	Experimental & Predicted	Multiple	Systematic information of plant, fungal, and micro-organism genomes [37]	982 proteins of 37 distinct interactions and 1859 PPIs
DIP [31]	2017	Secondary	Experimental	Multiple	Combinational information of DIP Nodes (Proteins), and DIP Edges (Interactions)	28850 Proteins, 834 Organisms, 81923Iinteractions, 82143 Distinct experiments, and 8234 Data sources
KEGG [32]	92.0	Primary	Experimental	Multiple	Origin and progression of cellular organisms [38]	PPI information of 6197 no. of organisms, 899 Disease-related elements, 380 Human gene variants, and 2337 Human diseases,11094 Drugs, and 2237 Drug groups
GenBank [33]	234.0	Primary	Experimental	Multiple	Combined information of GenBank at NCBI, DNA Data-Bank of Japan (DDBJ), and European Nucleotide Archive (ENA)	386197018538 Bases, and 216763706 Sequences
UniHI [34]	7.1	Secondary	Experimental & Predicted	Human	Complete drug-target dataset of Drug-Bank database [35]	16499 Genes, 158 Samples, 36023 Proteins, and 573995 Molecular interactions
GPS-Prot [36]	3.1.5	Secondary	Experimental	Human & HIV	HIV PPIs & Visualization of human protein sequences	Protein Interactions of 395501 Human, 8004 HIV-1, and 2291 HIV-1 Screen Hits
YPD [40]	6.0	Primary	Experimental	Saccharomyces Cerevisiae	Complete set of annotated data [39]	25000 lines of documentary annotation, 6021 entries, 2369 Known, 2421 Unknown, and 123 not characterized
3did [41]	2019_01	Secondary	Experimental	Multiple	Set of domain-domain interactions of 3D structure. PPI data to identify peptide-mediated interactions, and derived consensus motifs [42]	13499 Domain-domain interactions, 513184 Structures of domain-domain interactions, 812 Domain-motifs interactions, and 8223 Structures of domain-motif interactions
Pfam [43]	32.0	Primary	Experimental	Multiple	Collection of protein domain families. Each family is denoted by multiple sequence orientations and Hidden Markov Model (HMM)	17929 Total families, 1229 New families, 74.5% Pfamseq holds at least one Pfam domain, and 50.1% Residues fall inside Pfam domains
STRING [44]	11.0	Secondary	Experimental & Predicted	Human	Updated gene-set data and hierarchical clustering of linked network.	5090 Total organisms, 24584628 Proteins, and 3123056667 Interactions
BioGrid [45]	3.5.177	Primary	Experimental	Multiple	Collection of 66 organism's data from well recognized biomedical research with stress on central biological methods and human disorders [46]	1740143 Protein & genetic interactions, 28093 Chemical associations, 1350574 Non-redundant interactions, 12015 Non-redundant biological associations, and 28093 Raw biological associations

in the yeast dataset where both of these methods offer close and satisfactory accuracy measures.

A. PPI PREDICTION USING SVM WITH AUTO COVARIANCE DESCRIPTOR

In their work, Kumar H. *et al.* [47] employed [24], [25] for investigating PPI data in yeast. They used the dataset [29] to comprehend the overall scenario of interacting and non-interacting amino acids. By using the Auto Correlation Descriptor (ACR), [47] converted each amino acid descriptor, i.e., six in number, into uniform numerical strings.

The Auto Covariance (AC) or ACR is a function that measures the covariance of a process when pairing made between two points. The ACR descriptor was used for assigning six physicochemical descriptors to an individual amino acid residue of a protein sequence. The ACR is also used for comparing the autocorrelation measure between two protein sequences.

For an amino acid sequence say, AQGTALP, A was assigned numerical values of six descriptors, Q assigned with numerical values of six descriptors, and so on. After assigning numerical values, amino acid interactions computed, for the sake of simplicity, lengthy heterogeneous datasets converted into a homogeneous 180-dimensional vector (30 is the length of amino acid having 6 no. of descriptors) using ACR, where each vector represents a protein sequence. A pair of amino acid sequences say A and B can be concatenated in A+B or B+A way. Hence, [47] made cumulative concatenation between two such sequences to represent an interaction between them. The mathematical representation of the ACR is given in(4).

$$\mathrm{ACR}_{\mathrm{D,j}} = \frac{1}{\mathrm{L}-\mathrm{D}} \sum\nolimits_{\mathrm{pos}=1}^{\mathrm{L}-\mathrm{D}} (\mathrm{S}_{\mathrm{pos,j}} - \frac{1}{\mathrm{L}} \sum\nolimits_{\mathrm{pos}=1}^{\mathrm{L}} \mathrm{S}_{\mathrm{pos,j}}) \times (\mathrm{S}_{\mathrm{pos+D,j}} - \frac{1}{\mathrm{L}} \sum\nolimits_{\mathrm{pos}=1}^{\mathrm{L}} \mathrm{S}_{\mathrm{pos,j}} \quad (4)$$

In (4), ACR is autocorrelation value, j is jth descriptor, pos is position in sequence S, L is the length of the sequence, D is the distance between one descriptor to its neighbor.

The methodology adopted in [47] for PPI identification as follows:

- Interacting and non-interacting protein sequences obtained.
- Assigning numerical values to each of the six descriptors of amino acid.
- Heterogeneous length of numerical strings is converted to homogeneous data length using ACR.
- Cumulative concatenation of two protein sequences represents an interaction and constitutes the dataset for investigation.
- Dividing the complete datasets into two parts, i.e., training and test set.

- The training set is used to train classifiers [24], [25].
- The trained model is then subjected to the test validation.
- Finally, the result obtained and validated against protein data sequences of different organisms.

[47] referred to the work of [48]–[52] to create an N-dimensional hyper plane for classifying their data set using SVM. Initially, the classes labelled either 1 or 0. After classification, the entire dataset divided into two classes, i.e., class 1 of IP and class 0 of NIP. The authors used the SVM-RBF kernel because of its suitability in the dataset having class-conditional probability distribution close to the Gaussian distribution, which serves better accuracy in the binary classification. In [47], the data normalization was done using in (5).

$$x' = \frac{(x - min)\,(new_{max} - new_{min})}{(max - min) + new\_min} \tag{5}$$

In (5), x' is normalized value, x is the original descriptor value, max and min are maximum and minimum value of descriptor respectively.

After normalization, new-min becomes 0, and new-max becomes 1. Finally, the SVM classifier for 360 attributes, i.e., 180 no. of attributes each for sequence A and B employed. The training set was subjected to a 10-fold cross-validation technique, where the entire training data partitioned into ten equal-length sets, i.e., each set with 450 protein sequences and 360 no. of attributes. The model was then subjected to a test set of 1500 no. of protein sequences, out of which 1059 protein sequences were correctly classified, i.e., an accuracy of 70.6%.

An analytical tool, namely the Receiver Operating Curve (ROC) used for showing the classifier's performance by plotting the TP rate against the FP rate. For the RBF kernel parameter, $\Upsilon = 0.125$, the TP rate increases sharply with the FP rate until TP becomes 0.15, and after that, the TP rate is firmly linear to the FP rate.

Authors of [47] also used the back propagation algorithm to perceive the performance of the ANN model. The ANN model was trained until there is no variation in subsequent iterative values. The model was subjected to the training set, which generated a bi-classifier output with two class labels, i.e., 1 for interacting amino acid pairs and 0 for non-interacting pairs. The ANN model finally validated with the test dataset with an accuracy of 72.60%.

In a similar work, Guo Y. *et al.* [51] predicted PPI from protein sequences using SVM with AC. The authors used the AC variable to calculate the average relations between residues. The residue is the leftover material that remains after completion of a process or a set of processes. The calculation of the AC variable given in (6) is similar to ACR in (4).

$$AC_{v,d} = \frac{1}{l-v}\sum\nolimits_{p=1}^{l-v}\left(S_{p,d} - \frac{1}{l}\sum\nolimits_{p=1}^{l} S_{p,d}\right) \times \left(S_{(p+v),d} - \frac{1}{l}\sum\nolimits_{p=1}^{l} S_{p,d}\right) \tag{6}$$

In (6), d is a descriptor, p is the position of sequence S, l is the length of sequence S, v is the value of lag, and lag is the distance between one residue and its neighbouring residue. The calculation of no. of AC variables is given in (7).

$$N = D \times L \tag{7}$$

In (7), N is the no. of AC variables, D is the no. of descriptors, and L is the maximum lag.

The varied lengths of protein sequences resulted in vectors of uneven lengths. Therefore, to convert these vectors into uniform matrices, the authors used the Auto Cross-Covariance (ACC) method. The authors analysed vector sequences by referring to the work of Li *et al.* [53]. Finally, ACC created with two variables, i.e., CC for different descriptors and AC for similar descriptors. However, they considered only the AC variable to avoid creating large no. of variants. The details of the Dataset, Three-Level Strategy for NIP, and Performance Comparison of AC and ACC are specified below.

1) DATASET

Authors of [51] considered DIP database version DIP_20070219 [54] for collecting the PPI set of Saccharomyces Cerevisiae. To define the test subset, an Expression Profile Reliability (EPR) and the Paralogous Verification Method (PVM) [55] used. Initially, the subset consists of 5966 no. of protein pairs. However, to retain simplicity, the authors considered protein pairs having less than 50 amino acids. Finally, by using the cd-hit program [56], a data set of 5943 pairs derived. Since Non-Interacting pairs (NIP) are not exclusively available in the DIP [54] database, authors devised a 3 level strategy for creating NIP.

2) THREE-LEVEL STRATEGY FOR NIP

At first, the authors used the Prcp method to randomly generate NIP from a positive data set of [57] followed by deploying Psub, where subcellular localization of information is done using the Swiss-Port [58] which is a database of protein sequences rendering detail specification about protein sequence curation. Lastly, negative protein sequences are created using the Shufflet program [59], where adjustment of right-sided interacting pair sequences done for different values of k-let, k = 1,2,3.

The authors considered LIBSVM 2.84 [60] to employ [24]. The performance of [24] tested with five-fold cross-validation for a negative data set of 25 amino acids. The authors referred to the Jackknife test using two-fold cross-validation [61], [62] to optimize the RBF kernel parameters C and γ, respectively. The results achieved by [51] listed in Table 7 and shown in Fig. 5.

3) PERFORMANCE COMPARISON OF AC AND ACC

In [51], the AC converted into 420, i.e., 2 × 30 × 7 dimensional vectors and the ACC converted into 2940, i.e., 2 × 30 × 7 × 7 dimensional vectors. The negative dataset of Psub used for creating 5 test datasets, each with 30 no. of amino

TABLE 7. Result obtained using k-let; k = 1, 2, 3, Psub and Prcp.

Protein sequences (Negative Dataset)	Prediction Accuracy	Sensitivity	Precision
1-let	79.25%	79.29%	82.67%
2-let	77.30%	69.81%	85.14%
3-let	70.25%	60.74%	80.15%
Psub	86.23%	85.22%	87.83%
Prcp	58.42%	41.76%	62.64%

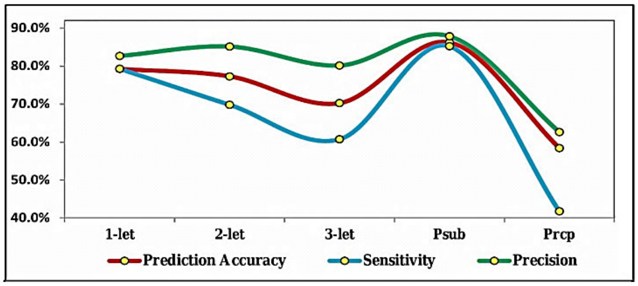


FIGURE 5. Comparison of results obtained using k-let; k = 1, 2, 3, Psub and Prcpprocess.

acids and 7 descriptors, characterizing each amino acid. The result achieved using AC and ACC for the SVM-RBF kernel parameter $\gamma = 0.0312$ listed in Table 8 and shown in Fig. 6.

TABLE 8. Comparison between ACC & AC for $\gamma = 0.0312$.

Method	Sensitivity	Precision	Accuracy
ACC	89.93	88.87	89.33
AC	87.30	87.82	87.36

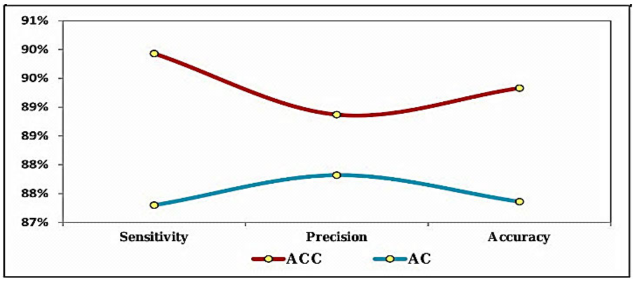


FIGURE 6. Performance of ACC & AC for $\gamma = 0.0312$.

It is evident from Table 8 and Fig. 6 that ACC performs better than AC. However, the authors considered the results of AC only to conclude the performance of the SVM method.

In [51], the performance analysis of the independent dataset was performed by considering 17491 pairs of yeast. By considering the residues with less than 50 amino acids, 11474 no. of pairs created of which 10108 no. of pairs predicted correctly, resulting in an accuracy of 88.09%. The authors generated the non-interacting test set pairs in the same locations by referring to the work of Ben-Hur *et al.* [63], where experimental testimony reveals that any non-co-localized negative protein pairs lead to the unfair and erroneous PPI prediction. Consequently, authors randomly selected a dataset of 8000 non-interactive cytoplasm and endoplasmic reticulum protein pairs where the classifier resulted in a moderate accuracy. However, without cytoplasm protein pairs, the classification accuracy reduced to 77%, and accuracy further reduced to 69% without endoplasmic reticulum protein pairs. In contrast, by conceding all 27204 no. of non-interactions, the overall prediction accuracy increased to 81.46%, and for non-redundant data set of 11474 no. of yeast data, the prediction accuracy further raised to 93.25%.

B. PPI PREDICTION USING SVM WITH CORRELATION COEFFICIENT

The Auto-correlation Descriptor (ACD) illustrates the correlation between two protein structures with the specific physicochemical property. The ACD is a topological descriptor that encrypts both physiochemical properties and molecular arrangement to the numerical vectors and represents it in uniform matrices, whereas the Correlation Coefficient (CC) quantifies the relationship strength between two variables. CC is used to transform the sizeable and heterogeneous protein sequence's physicochemical descriptions into a uniform length pattern. Therefore, both ACD and CC can be employed in the discrete arrangements of protein pairs' physicochemical attributes to reconstruct them into a consistent pattern. In this context, Shi *et al.* [64] employed CC with SVM classifier to predict the PPI using the yeast dataset with high accuracy.

1) DATASET

The authors considered the S. Cerevisiae positive data and employed the CC transformation to consider the neighboring effect of protein sequences and levels between protein pairs. The CC accepts 12 physiochemical properties of protein pairs and transformed it into a uniform pattern. They constituted the dataset using protein pairs of DIP, MIPS, and BIND databases. After removing protein pairs of B50 amino acids, 2,800 proteins, and 6,436 interactions retained with 829 proteins, 1025 interactions from MIPS and 736 proteins, 750 interactions for BIND databases. Finally, a total of 4365 proteins and 8211 interactions were considered to develop a positive dataset. Reference [64] generated the equal no. of negative dataset listed in Table 9.

2) METHODOLOGY

The authors considered 12 sequence-based physicochemical properties of 20 amino acids for this experiment. To normalize the physicochemical properties, the authors used the min-max normalization reprocessing method wherein CC is used to transform the sequence of protein pairs' physicochemical properties into a uniform shape. Therefore, CC for 12 physicochemical properties is used to measure the distance between

TABLE 9. Description of the negative dataset.

Serial No.	Name	Description
1	R-NEG	Non-interacting protein pairs collected randomly from the positive datasets
2	BS-NEG	Non-interacting pairs collected from Organelle DB database; wherein produced pairs form the discrete subcellular compartments in proportion to different subsets
3	IS-NEG	The non-interacting pairs in the same localization in Organelle DB database, not present in DIP, MIPS, and BIND databases.
4	GO-NEG	Negative protein pairs satisfying both the values of RSS Cellular Components and RSS Biological Processes lies between 0 and 0.4 with lower confidence by refereeing to the RSS similarity matrix of Wu et al. [65].

protein sequences using (8)-(10).

$$\mathrm{CC(s)} = \frac{\sum_{a=1}^{p-r} X_{a,b} \times \sum_{c=1}^{q-r} Y_{c,b}}{\sqrt{\sum_{a=1}^{p-r}\left(X_{a,b}\times X^{Z}_{a,b}\right)} \times \sqrt{\sum_{c=1}^{q-r}\left(Y_{c,b}\times Y^{Z}_{c,b}\right)}} \tag{8}$$

In (8), X, Y represent two protein pairs, respectively and s is the CC lag for d = 1, 2,. . . ,lg, where the maximum s is lg.

$$X_{a,b} = \left(M_{a,b} - \frac{1}{p}\sum_{a=1}^{p} M_{a,b}\right)\left(M_{a+r,b} - \frac{1}{p}\sum_{a=1}^{p} M_{a,b}\right) \tag{9}$$

$$Y_{c,b} = \left(N_{c,b} - \frac{1}{q}\sum_{c=1}^{q} N_{c,b}\right)\left(N_{a+r,b} - \frac{1}{q}\sum_{c=1}^{q} N_{c,b}\right) \tag{10}$$

In (9)-(10), a, c representing the positions of amino acid sequences M and N, b is 1 of 12 amino acids physicochemical properties, p and q are the lengths of sequences of amino acid M and N, respectively, d is the protein sequence distance of two different residues. In (10), s is the CC lag, d = 1, 2,. . . ,lg, wherein the maximum s is lg.

The CC variables calculated 12 descriptors and 12 × lg descriptor values. After generating the protein sequence vector space with 12 × lg dimension, using the CC transformation, the new vector set divided into k subsets and replications of k times with k-fold cross-validation, i.e., each kth subset as the test set and remaining k - 1 subset as the training sets. Authors considered LIBSVM [60] software for the SVM method with RBF kernel and measured the classification performance and the prediction performance (PP) is given in (11).

$$\mathrm{PP}\left(V \geq V_o\right) = \frac{\#\left(V \geq V_o\right)}{\#\left(R\right)} \tag{11}$$

In (11), v is the validated value of SN, PE, MCC, and ACC, V_O is randomly generated values from observation datasets, and R is randomly generated observation datasets.

3) RESULT

From the final positive and negative dataset of 2,050 protein pairs, the authors randomly selected 50%, i.e., 1,025 pairs for the training set, and the remaining 50%, i.e., 1,025 for the test set, which was trained by the SVM method with 5-fold cross-validation for five iterations repeatedly resulting a total of 3 positive datasets of S. Cerevisiae and four negative datasets to measure the performance of the SVM method. From these four negative datasets, the GO-NEG negative dataset outperforms the other dataset and MIPS core positive with GO-NEG negative dataset, which resulted in height values of 86.86% sensitivity, 89.52% Precision, 75.98% Mathew's Correlation Coefficient (MCC), and 87.94% accuracy averaging more than 5.2%, 2.5%, and 7.5% of R-NEG, BS-NEG, and IS-NEG, respectively. Additionally, the authors compared the performance with CC and ACD transformation of protein sequence for 5 test sets using the SVM model using MIPS core positive with GO-NEG negative dataset for 24 amino acids shown in Table 10 and shown in Fig. 7. From Table 10, it is clear that the SVM-CC transformation performed better than AC.

TABLE 10. SVM performance for MIPS corewith GO-NEG dataset.

	Sensitivity	Precision	MCC	Accuracy
SVM-CC	86.86	89.52	75.98	87.94
SVM-ACD	85.11	83.94	69.40	84.70

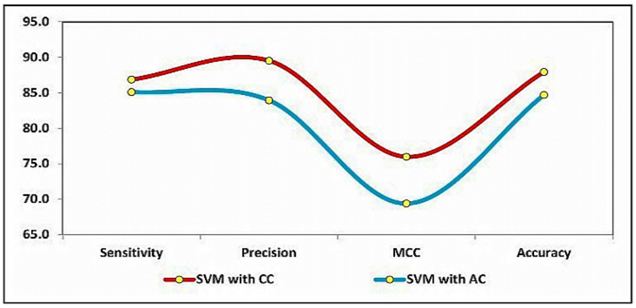


FIGURE 7. SVM-CC and SVM-ACD performance.

The SVM method also delivers better performance with Boosting [66] andLasso [67] methods for the Helicobacter pylori [68] dataset. In another work, [69] employed ACD with Rotation Forest as both ACD and CC are equally prudent in the discrete arrangements of protein pairs' physicochemical attributes reconstructed to a uniform pattern, [69] considered Rotation Forest with ACD on Saccharomyces Cerevisiae and Helicobacter Pylori data from the DIP database and compared the performance with Guo *et al.* [51], wherein SVM-AC offered an adequate prediction accuracy of 93.97%. In another work, Ma *et al.* [70] employed six classifiers, including K-Nearest Neighbour (KNN), ANN, RFM, Naive Bayes, Logistic Regression, and SVM methods for sequence-based prediction on Helicobacter pylori and Human protein pairs of DIP database. Authors implemented 5-fold, 8-fold, and 10-fold cross-validation for these six classifier models and observed that the SVM method outperforms the other five models for the H. pylori protein pairs with a prediction accuracy of 72.79% and human protein pairs with an accuracy of 83.88%.

C. PPI PREDICTION USING DOMAIN PROPERTY AND TWO-CLASS SVM

In their work, Chatterjee *et al.* [71] adopted the concept of domain-domain affinity and the two-class SVM to extract features from the dataset [54] and to predict PPI, respectively. Consequently, the approach employed in [71] is shown in Fig. 8.

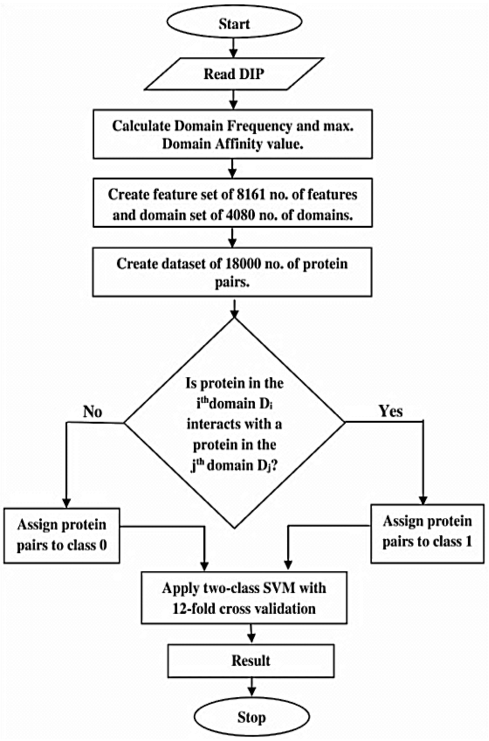

Start
Read DIP
Calculate Domain Frequency and max. Domain Affinity value.
Create feature set of 8161 no. of features and domain set of 4080 no. of domains.
Create dataset of 18000 no. of protein pairs.
Is protein in the i^th domain Di interacts with a protein in the j^th domain Dj?
No
Yes
Assign protein pairs to class 0
Assign protein pairs to class 1
Apply two-class SVM with 12-fold cross validation
Result
Stop

FIGURE 8. Methodology flowchart.

The detailing of Domain Frequency and Domain Affinity exercised by [71], along with the Dataset, Results, and Comparative Analysis stated underneath.

1) DOMAIN FREQUENCY AND DOMAIN AFFINITY

A domain is a functional and structural element for which a specific sequence of the PPI pattern preserved. Elementarily the PPI data is disintegrated into physical associations between constituting domains of the respective proteins. In this context, the authors of [71] used two fundamental domain characteristics, Domain Frequency (DF), and Domain Affinity (DA), to extract features from [54] shown in Fig. 9.

In [71] referred to the profile method [72], which elaborates innovative pieces of data about domain interactions, whereas, a list of domain affinities characterizes protein pairs. To construct the feature set, the authors considered DF and DA values represented as V^d and AF, respectively, for each protein pair. For a PPI pair, the vector V_i^d is the frequency of the i^{th} domain in a range [0, 1] given in(12).

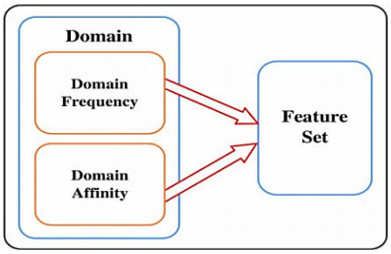


FIGURE 9. Domain based extraction of features.

$$V_i^d = \begin{cases} \text{Value of DF,} & \text{if pair is in i}^{th}\text{domain} \\ 0, & \text{Otherwise} \end{cases} \quad (12)$$

Therefore, a PPI is a result of the interaction between the corresponding domain pairs, where AF used to identify all such interactions. The AF estimation for two interacting domain pairs, say, P_i and P_j is given in(13).

$$AF = \text{Max.}\left(\frac{\text{Affinity}(P_i, P_j)}{100}\right) \quad (13)$$

The authors used all possible combinations of single and multiple domain protein pair sequences to extract the feature set. Finally, by using DF and DA, a set of 8161 no. of features from 4080 no. of unique Pfam [43] domains extracted. [43] is a big collection of protein relations and describes the Hidden Markov Model (HMM) of protein domains.

2) DATASET

The authors referred to [54] for assembling two types of PPI information for their analysis. They considered the stable data of protein-protein complexes as well as temporary complexes where proteins momentarily bind with each other to attain a specific purpose. Initially, a total of 9000 protein pairs (binary strings) were collected; however, due to the non-availability of NIP in [54], authors artificially created a random set of 9000 NIP using the Exhaustive Search Method (ESM) [73].

The ESM is a brute force search technique for identifying all possible solutions for a given problem and satisfying each solution against the problem statement. For a given function f (p) and x no. of intermediate points, namely $p_1, p_2, p_3, \ldots, p_x$, in a given range [m, n], [73] is shown in Fig. 10 along with an algorithmic description as follows.

Finally, a dataset of 18000 pairs created following which the entire dataset was equally partitioned into two classes, namely, class 0 of 9000 NIP and class 1 of 9000 IP.

3) RESULT

For deploying SVM, authors of [71] referred to the SVM-light code developed by Joachims [74]. Reference [71] considered the linear kernel, the polynomial kernel of degree 2, and the RBF kernel with $\gamma = 0.00123$. A twelve-fold class validation method employed for the two-class SVM method for a training set size of 1500 no. of samples,

```
Algorithm 1 ESM (f(p), m, n)
  Input: Function f(p) and x no. of intermediate points between m and n.
  Output: Distinct x no. of values between m and n.
  Step 1: set $p_1 = m$ and $\Delta p = (n - m)/x$
          // $p_2 = p_1 + \Delta p$, $p_3 = p_2 + \Delta p$ and so on....
  Step 2: if $f(p_1) \geq f(p_2) \leq f(p_3)$ then
          // minimum point in $[p_1, p_3]$
      stop
      else
      $p_1 = p_2$
        $p_2 = p_3$
          $p_3 = p_2 + \Delta p$
        end if
  Step 3: if $p_3 \leq x$ then
        go to 2
        else
        minimum value does not exist in [m, n]
  end if
```

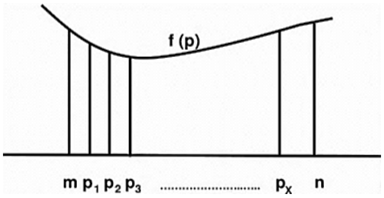


FIGURE 10. Exhaustive search method.

i.e., 8.33% of the total sample having 750 positives and 750 negative protein pairs. The result obtained in [71] using different kernel functions are listed in Table 11 and shown in Fig. 11.

TABLE 11. Result of different kernel functions.

Kernel	Accuracy	Precision /Specificity	Recall / Sensitivity
RBF	86%	95.35%	75.65%
Linear	86%	95.24%	75.71%
Polynomial	69%	84.96%	46.00%

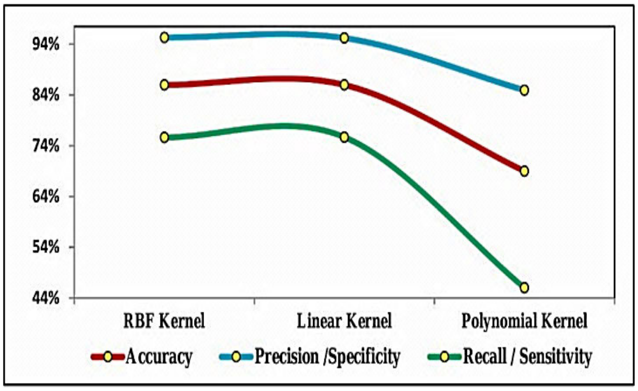


FIGURE 11. Result comparison for different kernels.

It is evident from Fig. 11 that RBF and linear kernels produced a reasonable classification accuracy of 86%.

4) COMPARATIVE ANALYSIS

Authors of [71] compared their outcome with the results achieved by Chen *et al.* [75], Han *et al.* [76], and Alashwal *et al.* [77]. Reference [75] employed a domain-based approach and achieved a sensitivity of 79.3% and a precision of 62.8%, whereas, [76] used the ranking method and reached a sensitivity of 77% and specificity of 95 %, the authors of [77] employed [47] and achieved a sensitivity of 77.4% and a specificity of 83.9%. The authors also compared their results with the outcomes achieved by Zaki [78] and Kim *et al.* [79]. Reference [78] analysed the information of inter-domain linker regions and found 60% sensitivity and 70.26% specificity, whereas by employing the Potentially Interacting Domain (PID) method, i.e., a domain-based algorithm, used for evaluating the interaction probability between protein pair domains, [79] achieved 50% sensitivity and 98% specificity. A comparison of specificity and sensitivity achieved by [71] with others is listed in Table 12 and shown in Fig. 12.

TABLE 12. Comparison between Chatterjee *et al.* with Other authors.

Authors	Method	Sensitivity	Specificity
Chatterjee P.et al. [71]	Two-class SVM	75.65%	95.35%
Chen X.W.et al. [75]	Domain-based	79.30%	62.80%
Han D.S.et al. [76]	Ranking method	77.00%	95.00%
Alshawl H.et al. [77]	BayesianClassification	77.40%	83.90%
Zaki N. [78]	Inter domain linker regions	60.00%	70.26%
Kim W.K.et al. [79]	PID	50.00%	98.00%

D. PPI IDENTIFICATION USING NORMALIZED POLYPEPTIDES AND SVM

In their work Romero-Molina S. *et al.* [80] accumulated the information of amino acid sequences and applied a mathematical tool to represent a normalized form of polypeptides, following which the SVM method is employed to predict protein pair interactions. Moreover, they also developed a PPI-detect predictor to detect peptides that bind better than EPI-X4, which is an endogenous peptide inhibitor of CXCR4 and G-protein-coupled receptor [61], [81], [82]. [80] referred to the work of Chou [83] to adopt five-step rules

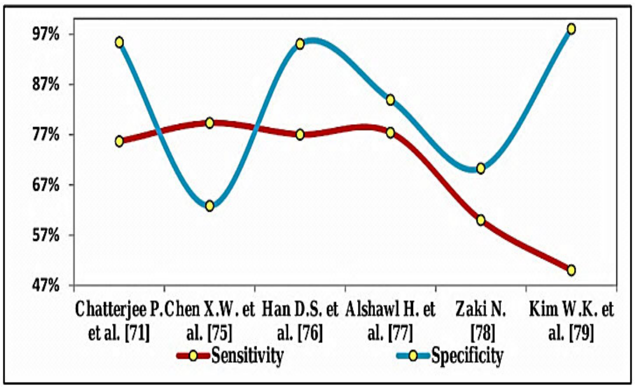


FIGURE 12. **Comparison between Chatterjee P. *et al.* with other authors.**

for producing a sequence-based mathematical predictor. The methodology of [80] as follows:

1. Training and testing the predictor by choosing an appropriate benchmark dataset.
2. Representing the biological patterns with a mathematical expression that imitates their association with the target application.
3. Designing a powerful engine to execute the prediction process.
4. Implementing cross-validation tests for calculating the correctness of the predictor.
5. Developing a publicly accessible predictor through network server architecture.

The specifications of Dataset, Validation, and Performance Comparison with other Predictors and Significance of PPI-Detect are mentioned below.

1) DATASET

The authors referred to three databases to construct the experimental dataset, i.e., [41] for obtaining 3D interacting domains, iPfam [84] for obtaining the domain interactions and protein families, and [27]. A total of 9326 no. of domain pairs obtained from [41], 9516 no. of pairs from [84], and 2666 no. of pairs from [27]. Authors considered the individual domain pairs with distinct elements present in both positive and negative sets. Finally, 1922 no. of interacting domain pairs, and 2405 no. of non-interacting domain pairs constituting 1922 + 2405 = 4327 no. of pairs obtained for the analysis. These 4327 no. of pairs are subdivided for training and testing their model. A training set of 3491 no. of PPI pairs, i.e., 1613 positives + 1878 negatives constituted and a test set of 836 no. of pairs, i.e., 309 positives + 527 negatives created. To evaluate the performance of their proposed model, they partitioned the test set into three classes as follows:

1. Very hard: The domain pairs which remain absent in the training set. This set consists of 103 no. of domain pairs, i.e., 57 positives + 46 negatives.
2. Medium-hard: The domain pairs in which at most one domain is present in the training set. It consists of 307 no. of domain pairs, i.e., 102 positives + 205 negatives.
3. Easy: The domain pairs in which both domains present in the training set. It consists of 426 no. of domain pairs, i.e., 150 positives + 276 negatives.

The authors of [80] referred to their earlier work to employ a method, namely ProtDCal [85], which is a protein-feature generation system tool. The purpose of using ProtDCal is to encode protein orders and structures. Level-wise explanations of encoding of distinct proteins using [85] are as follows:

- **Level 1**

The physicochemical and structural characteristics of the amino acid set fed to the ProtDCal tool to generate a residue-based feature matrix.

- **Level 2**

A noteworthy feature of [85] is that it can only ply with individual amino acid sequences. Therefore, to generate a pairwise descriptor from a single-chain descriptor of amino acid sequences, authors of [80] developed a strategy as follows:

Let X and Y be a pair of amino acid sequences, then the concatenation of the pair, i.e., XY or YX representing block co-polymers made based on the inequality, 2X + 2Y > XY+ YX. Subsequently, the pairwise descriptor represented as C_{X-Y} is given in (14).

$$C_{X-Y} = C_{XY} + C_{YX} - 2C_X - 2C_Y \tag{14}$$

In (14), C_X, C_Y, C_{XY}, and C_{YX} represents the single-chain descriptor values for sequences X, Y, XY, and YX, respectively.

The above approach can also be employed to generate a vicinity-modified residue feature matrix from the residue-wise feature matrix using Electro-topological State (ESO) operator [86] given in (15).

$$M_{ESO} = M_a - \sum\nolimits_{b\neq a}^{R} \frac{M_b - M_a}{(b-a)^2} \tag{15}$$

In (15), M_{ESO} represents the vicinity-modified index using the E-State operator, M_a is the value of the M^{th} index of the amino acid sequence for the residue a, M_b is the value of the M^{th} index of the amino acid sequence for the residue b where a and b both belong to the residue set R.

- **Level 3**

Based on the residue properties, the divide-and-conquer approach is employed to split the vicinity-modified residue feature matrix into multiple no. of group-based matrices.

- **Level 4**

Finally, a protein-feature matrix of dimension D ×F inferred from the vicinity-modified residue feature matrix where D is the no. of proteins, F is the no. of features. The dimension F extended as $F = A \times C \times O$, where A is the no. of amino acid properties, C is the no. of grouping criteria, and O is the no. of aggregation operators of amino acids.

Initially, a total of 3,248 features obtained for each pair of proteins, to reduce the dimension of such an extensive feature set, the authors adopted the Modeling Protocol as follows:

1. A scoring mechanism was employed by the authors using Weka 3.7.11 [87] package for a content threshold of 5%. Consequently, by eliminating the features that

distinguish interacting domains from non-interacting domains in the training set, the feature set was reduced to 326 no. of features.

2. The DCluster tool [85] used in [80] to extract the required components from each cluster, thereby eliminated redundancy in the dataset. For a threshold value 0.95, the feature set was further reduced to 322 no. of features.
3. By employing the WrapperSubsetEval method of [87] and the Genetic Algorithm (GA) [88], the authors identified and estimated multiple sets of attributes of the features. Lastly, by considering a population size of 20 no. of samples with a mutation and crossover probabilities 0.033 and 0.6 respectively, the SVM method was applied with the RBF and polynomial kernel following which the grid-search method Hsu C.W. *et al.* [89] used with five-fold cross-validation. However, for simplification, [80] considered the linear kernel with C = 11.3, to obtain an optimal set of 19 no. of features.

2) VALIDATION

The authors applied the Precision-Recall Curve (PRC) with a 10-fold cross-validation technique on the entire dataset to determine the performance of the PPI-detect method. They obtained a precision of 90% and a sensitivity of 30%. Though, the PRC analysis reveals that 50% of precision classified with 90% of sensitivity, whereas 50% of sensitivity created with 78% of precision. Therefore, it implies that the PPI data identified by the ten-fold cross-validation technique was biased. To overcome the bias factor, [80] split the test set into three classes, i.e., easy, mid-hard, and very hard. The precision and sensitivity values for these three test sets listed in Table 13.

TABLE 13. Precision and sensitivity of test sets.

Test Subset	Precision	Sensitivity
Easy	90%	70%
Mid-hard	90%	45%
Very hard	90%	45%

3) PERFORMANCE COMPARISON OF THE PREDICTORS

For a probability threshold value of 0.5 on Mid-hard and Very-hard test groups, the authors compared the proposed PPI-predictor with other widely used predictors, i.e., Pred-PPI [51], PIPE [90], and SPPS [91]. PIPE is a sequence alignment-based predictor whereas Pred-PPI and SPPS are SVM predictors. Comparative results obtained using different predictors listed in Table 14 and shown in Fig. 13.

It is evident from Fig. 13 that the PIPE prediction provides the highest precision of 0.76 and Pred-PPI give maximum sensitivity of 0.88, whereas, the PPI-Detect predictor performs reasonably with an accuracy of 0.66.

TABLE 14. Performance comparison with PPI-Detect.

Method	Pred-PPI[51]	PIPE [90]	SPPS [91]	PPI-Detect [80]
Precision	0.40	0.76	0.51	0.55
Sensitivity	0.88	0.10	0.24	0.65
Accuracy	0.44	0.64	0.62	0.66

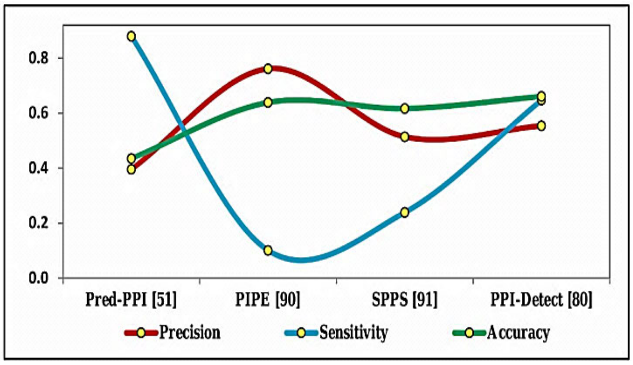


FIGURE 13. Performance comparison between PPI-Detect with other predictors.

4) SIGNIFICANCE OF PPI-DETECT

The authors applied the PPI-Detect predictor to predict and identify the working derivatives of EPI-X4 interactions. According to Zirafi O. *et al.* [82], EPI-X4 is an endogenous antagonistic ligand of the CXC Chemokine Receptor 4 (CXCR4) which is a G-protein-linked receptor [61], [81]. By supplying a total no. of 35 peptides to the PPI-Detect predictor, the precision and accuracy values noted. A summary of precision and accuracy achieved in identifying active EPI-X4 derivatives based on the predicted interaction with four fragments of CXCR4 is listed in Table 15 and shown in Fig. 14.

TABLE 15. Performance of PPI-Detect on EPI-X4.

Fragment	Residues	Precision	Accuracy
FRAGMENT A	25–45	0.52	0.63
FRAGMENT B	87–121	0.70	0.71
FRAGMENT C	164–205	0.50	0.60
FRAGMENT D	252–292	0.20	0.51

From the results listed in Table 15, it is quite apparent that fragment B furnishes a high precision rate of 70%. Therefore, based on the result, it can be concluded that the EPI-X4 may bind more firmly with the CXCR4. The authors also considered three small derivatives of CXCR4, namely, JM130, JM133, and JM135. The prediction result of these three derivatives revealed that the JM133 is roughly 3 times active than the EPI-X4.

IV. RESEARCH DIRECTION AND CHALLENGES

This section broadly incorporates a summary of PPI classification results reviewed in this publication, along with the opportunities and challenges of the PPI study.

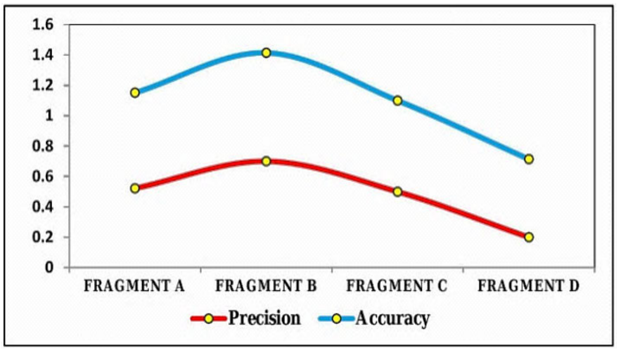


FIGURE 14. Realization of PPI-Detect on EPI-X4.

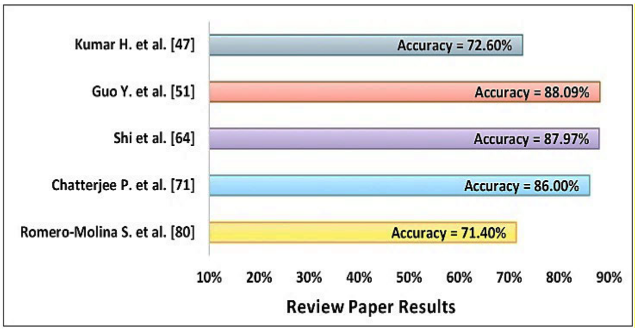


FIGURE 15. Performance summary of review papers.

TABLE 16. Summary results of review paper performance.

Author	Method	Dataset	Results
Kumar H. et al. [47]	SVM & ANN	KUPS	Accuracy = 72.60%
Guo Y. et al. [51]	SVM	DIP	Accuracy = 88.09%
Shi et al. [64]	SVM	DIP, MIPS, and BIND	Accuracy = 87.94%
Chatterjee P. et al. [71]	SVM	DIP	Accuracy = 86.00%
Romero-Molina S. et al. [80]	SVM	3did, iPfam and Negatome	Accuracy = 71.40%

In [47], the SVM method used for a test dataset of 1500 protein sequences and correctly classified data with an accuracy of 70.6 %. They also achieved an accuracy of 72.6% for the ANN method. In a similar work, authors of [51] combined SVM with the AC concept and produced a high accuracy in PPI classification. Reference [51] devised a three-level strategy using a k-let approach for NIP and achieved an accuracy of 88%. They also compared the accuracy results of SVM-AC, SVM-ACC and concluded that for their dataset, ACC outperforms the AC method.

Lately, Shen *et al.* [57] proposed the SVM-based prediction model with a conjoint triad feature to predict PPI networks with reasonable accuracy. However, their model can predict interaction networks in human PPIs for continuous amino acid chains without considering neighbouring effects. These limitations were addressed by [51, 64], wherein [64] used the SVM-CC method to predict yeast-PPI with competent results.

In the work of [71], the concept of a domain-domain affinity was used for selecting the feature set. For 9000 no. of IP and 9000 no. of artificially curated NIP, they achieved an accuracy of 86 % for SVM-RBF.

In [80], a combination of the SVM method and normalized polypeptides was used for PPI classification. The authors considered an array of databases to create an initial dataset of 21508 no. of domain pairs, and after careful filtration, 1922 no. of IP and 2405 no. of NIP considered for the investigation. The PPI-Predictive model of [80] performed convincingly compared to [51], [90], [91] predictors. A summary of the results reviewed in this publication is listed in Table 16 and shown in Fig. 15.

A. RESEARCH DIRECTION

In the medical investigation, the use of [24], [25] is highly significant. In this context, Lo *et al.* [92] employed the [25] for computer-simulated analysis of mammography using an optimal no. of feature inputs, and the result revealed that for four no. of feature inputs the [25] knocked the results of conventional medical radiology. Polat *et al.* [93] employed a Least-Square SVM (LS-SVM) method to analyze breast cancer data and delivered symptomatic accuracy of 98.53%, whereas, in their work, Enyinnaya [94] combined SVM and Sequential Minimal Optimization (SMO) algorithm, i.e., SVM - SMO, to identify cancer-causing proteins in PPI. In contrast, Chuang *et al.* [95] implemented the classification of PPI network for determining biomarkers as subnets, and their result provided tumor sequence pathways correctly. The authors of [96] suggested that [25] can be employed to diagnose PPI in Human-Mycobacterium tuberculosis. In this approach, the intra-race training of [25] using a combination of human PPI and Bacillus Anthracis data of different species was done for inter-race forecasting, resulting in a binary classifier that predicted traces of Bacillus Anthracis in human with moderate accuracy of 89.0%. Recently, Dey *et al.* [97] employed various ML models, including SVM, to predict interactions between SARS-CoV2 and human protein pairs, wherein the SVM method performed adequately for RBF and polynomial kernels, with the accuracy of 69.67% and 68.03%, respectively. However, [97] proposed an ensemble technique that outperformed the other models with an accuracy of 72.33%. Lastly, we present a summary of the contributions and limitations of the publications reviewed, listed in Table 17.

B. CHALLENGES

The employability and effectiveness of the SVM method are often limited by the selection of the inappropriate kernel functions, which leads to erroneous discrimination between solvent and insolvent problems, thereby resulting in inferior solutions. An appropriate selection of SVM-Kernel aids in minimizing the cost overhead associated with the transformation of data from linearly non-separable to separable ones thereby reduces human interventions. Therefore, optimally shaped SVM parameters along with appropriate kernel

TABLE 17. Summary of research contribution and limitations of review paper.

Author	Contribution	Limitations
Kumar H. et al. [47]	SVM and ANN classifiers used for PPI prediction offering more efficient results than existing sequence-based methods. The proposed model also validated for the dataset of Plasmodium Falciparum and Stem cell.	An extensive comparison between other ML methods not considered.
Guo Y. et al. [51]	The SVM, AC, and ACC methods combined to predict PPI. Authors constructed nine different models, including nine different lags in Psub negative data-set with five negative training data-sets Psub, Prcp, 1-let, 2-let, and 3-let, to predict PPI.	Only one organism, i.e., Saccharomyces Cerevisiae, used to demonstrate the effectiveness of SVM in PPI prediction.
Shi. et al. [64]	PPI prediction based on twelve sequence-based physicochemical properties of 20 amino acids of S. Cerevisiae. The authors employed SVM-CC to predict PPI, which outperformed SVM-AC.	The model's performance can be tested for diverse datasets, both empirical and actual.
Chatterjee P. et al. [71]	PPI prediction accomplished with SVM classification. The authors used the concept of Domain-domain interaction frequency and Domain affinity to build the features set. Their model achieved reasonable accuracy compared to other PPI prediction methods with similar domain information.	The authors created manually curated random NIP, which may not exist in reality. Additionally, different types of protein features such as solvent accessibility, subcellular localization, and hydrophobicity can be considered to enhance the prediction performance.
Romero-Molina S. et al. [80]	The authors developed a ProtDCal tool to transforms amino acid sequences into vectors suitable for ML algorithms. They proposed the PPI predictor method, namely PPI-Detect, which outperformed conventional predictors. The said predictor applied to the derivatives of EPI-X4 to establish the anti-CXCR4 footprints effectively.	The PPI-Detector applied only on a particular G-protein-coupled receptor CXCR4 without considering other receptors.

selection help in inferring better specimens with novel solutions in spite of the presence of bias in the input. However, a significant drawback of the SVM method is the transparency problem, which becomes substantial for higher-dimensional datasets, where the classification result may not always sketch a parametric function of PPI characteristics.

The quality of SVM-based solutions is proportional to the quality of the data. A comprehensive resource of PPI databases well supplements the expansion of PPI research through PPP (Private-Public Partnerships) model. Though, widespread obstacles in access, treatment, and synthesis of these databases block their fullest use. PPI databases mostly revolve around a specific type of test data such as Nucleotide-sequence or Protein-sequence data. However, in practice, a majority of these databases represent the data oddly and utilize a variety of encoding mechanisms to represent protein-related information, i.e., attribute names, units of measure, which often leads to an Interoperability Problem. A solution to the interoperability problem is database integration, which can be accomplished by adopting the concept of Data Warehousing. In the Warehousing approach, a large no. of databases connected in response to a query, i.e., Sequence-Retrieval System (SRS), where databases are handled as text files and indexing of these files, in turn, are based on keywords and PPI attributes.

V. CONCLUSION

The modern time is witnessing an outbreak of high-quality genomic information that warrants the use of sound methods such as machine learning to address multifaceted problems in PPI study. However, the exponential growth in PPI information makes the job of database curators tedious in storing pieces of protein data suitably wherein the productive use of the query response system helps in delivering the fitting PPI information for ML-based classifications. The selection of ML-methods for investigating PPI becomes more effective than traditional exhaustive and costlier approaches because ML-methods offer robust solutions. In this context, examining PPI through the lenses of machine classifiers such as SVM and ANN delivers prudent outcomes because of their abilities in automating the learning process without being programmed explicitly. Though, machine classifiers often exhibit biased results that reduce the nobility of the solution, i.e., the unavailability of NIP information in the DIP database disproportionate the presence of IP and NIP, leading to biased outcomes, which authorizes the formation of rational approaches to create NIP artificially. The coherent use of statistical descriptors, including CC, ACR, and ACD, offers compelling sequence-based predictions with the SVM method.However, protein functions are instrumental in PPI analysis, where the usefulness of classification approaches other than SVM and ANN cannot be overlooked. Consequently, an assemblage of classification methods such as Bayesian, nearest-neighbor, and $\chi 2$ are useful in determining protein functions from a fixed no. of functional categories. However, the Bayesian method produces more favorable outcomes in terms of sensitivity measure than the nearest-neighbor and $\chi 2$ methods. Selecting features using the concept of domain opens new avenues in the feature extraction process. Lately, SVM's hybrid approach, coupled with normalized polypeptides, was used to develop a PPI-detector model, which outperformed other conventional predictor's accuracy. The ML-classification outcomes are significant in presenting novel acumens into the regulative devices to spot biomarkers essential for the prognosis of various diseases such as leukemia. Despite the precision offered by the SVM method in classification, most of the outcomes still need empirical validation, thereby offering a broad spectrum of research openings.

REFERENCES

[1] J. F. Xia, S. L. Wang, and Y. K. Lei, "Computational methods for the prediction of protein-protein interactions," *Protein Peptide Lett.*, vol. 17, no. 9, pp. 1069–1078, 2010.

[2] J. Feng, R. Jiang, and T. Jiang, "A max-flow based approach to the identification of protein complexes using protein interaction and microarray data," in *Proc. Comput. Syst. Bioinf.*, Stanford, CA, USA, Aug. 2008, pp. 1–12.

[3] X. L. Li, S. H. Tan, C. S. Foo, and N. See-Kiong, "Interaction graph mining for protein complexes using local clique merging," *Genome Informat.*, vol. 16, no. 2, pp. 260–269, 2005.

[4] A. D. King, N. Przulj, and I. Jurisica, "Protein complex prediction via cost-based clustering," *Bioinformatics*, vol. 20, no. 17, pp. 3013–3020, Nov. 2004.

[5] D. Bu, "Topological structure analysis of the protein-protein interaction network in budding yeast," *Nucleic Acids Res.*, vol. 31, no. 9, pp. 2443–2450, May 2003.

[6] S. V. Rao, K. Srinivas, G. N. Sujini, and G. N. S. Kumar, "Review article: Protein-protein interaction detection: Methods and analysis," *Int. J. Proteomics*, vol. 2014, pp. 1–12, Feb. 2014.

[7] P. Artursson, K. Palm, and K. Luthman, "Caco-2 monolayers in experimental and theoretical predictions of drug transport," *Adv. Drug Del. Rev.*, vol. 46, nos. 1–3, pp. 27–43, 2001.

[8] D. Relman, "Detection and identification of previously unrecognized microbial pathogens," *Emerg. Infectious Diseases*, vol. 4, no. 3, pp. 382–389, Sep. 1998.

[9] S. Klein, "The use of biorelevant dissolution media to forecast the *in vivo* performance of a drug," *AAPS J.*, vol. 12, no. 3, pp. 397–406, Sep. 2010.

[10] E. Roberts, A. Magis, J. O. Ortiz, W. Baumeister, and Z. Luthey-Schulten, "Noise contributions in an inducible genetic switch: A whole-cell simulation study," *PLOS Comput. Biol.*, vol. 7, no. 3, pp. 1–21, 2011.

[11] I. S. Moreira, P. A. Fernandes, and M. J. Ramos, "Hot spots—A review of the protein-protein interface determinant amino-acid residues," *Proteins, Struct., Function, Bioinf.*, vol. 68, no. 4, pp. 803–812, Jun. 2007.

[12] T. Clackson and J. Wells, "A hot spot of binding energy in a hormone-receptor interface," *Science*, vol. 267, no. 5196, pp. 383–386, Jan. 1995.

[13] J. Kenneth Morrow and S. Zhang, "Computational prediction of protein hot spot residues," *Current Pharmaceutical Design*, vol. 18, no. 9, pp. 1255–1265, Mar. 2012.

[14] J. A. Wells, "Systematic mutational analyses of protein-protein interfaces," in *Methods Enzymology*, vol. 202. Cambridge, MA, USA: Academic, 1991, pp. 390–411.

[15] K. S. Thorn and A. A. Bogan, "ASEdb: A database of alanine mutations and their effects on the free energy of binding in protein interactions," *Bioinformatics*, vol. 17, no. 3, pp. 284–285, Mar. 2001.

[16] J. Xia, Z. Yue, Y. Di, X. Zhu, and C.-H. Zheng, "Predicting hot spots in protein interfaces based on protrusion index, pseudo hydrophobicity and electron-ion interaction pseudopotential features," *Oncotarget*, vol. 7, no. 14, pp. 18065–18075, Apr. 2016.

[17] S. Lise, C. Archambeau, M. Pontil, and D. T. Jones, "Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods," *BMC Bioinf.*, vol. 10, no. 1, pp. 1–17, Dec. 2009.

[18] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, 2000.

[19] N. Tuncbag, O. Keskin, and A. Gursoy, "HotPoint: Hot spot prediction server for protein interfaces," *Nucleic Acids Res.*, vol. 38, pp. W402–W406, Jul. 2010.

[20] T. B. Fischer, K. V. Arunachalam, D. Bailey, V. Mangual, S. Bakhru, R. Russo, D. Huang, M. Paczkowski, V. Lalchandani, C. Ramachandra, B. Ellison, S. Galer, J. Shapley, E. Fuentes, and J. Tsai, "The binding interface database (BID): A compilation of amino acid hot spots in protein interfaces," *Bioinformatics*, vol. 19, no. 11, pp. 1453–1454, Jul. 2003.

[21] S. Lise, D. Buchan, M. Pontil, and D. T. Jones, "Predictions of hot spot residues at protein-protein interfaces using support vector machines," *PLoS ONE*, vol. 6, no. 2, pp. 1–7, 2011.

[22] Y. Qiao, Y. Xiong, H. Gao, X. Zhu, and P. Chen, "Protein-protein interface hot spots prediction based on a hybrid feature selection strategy," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–16, Dec. 2018.

[23] C. Von Mering, R. Krause, and B. Snel, "Comparative assessment of large-scale data sets of protein-protein interactions," *Nature*, vol. 417, pp. 399–403, May 2002.

[24] V. Vapnik, "Pattern recognition using generalized portrait method," *Automat. Remote Control*, vol. 24, no. 6, pp. 774–780, Jan. 1963.

[25] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[26] A. C. Eschenlauer and W. S. Reznikoff, "Escherichia coli catabolite gene activator protein mutants defective in positive control of LAC operon transcription.," *J. Bacteriology*, vol. 173, no. 16, pp. 5024–5029, 1991.

[27] P. Blohm, G. Frishman, P. Smialowski, F. Goebels, B. Wachinger, A. Ruepp, and D. Frishman, "Negatome 2.0: A database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D396–D400, Jan. 2014.

[28] B. J. Breitkreutz, C. Stark, and M. Tyers, "The GRID: The general repository for interaction datasets," *Genome Biol.*, vol. 4, no. 3, pp. 1–3, 2003.

[29] X.-W. Chen, J. C. Jeong, and P. Dermyer, "KUPS: Constructing datasets of interacting and non-interacting protein pairs with associated attributions," *Nucleic Acids Res.*, vol. 39, pp. D750–D754, Jan. 2011.

[30] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stumpflen, H.-W. Mewes, A. Ruepp, and D. Frishman, "The MIPS mammalian protein-protein interaction database," *Bioinformatics*, vol. 21, no. 6, pp. 832–834, Mar. 2005.

[31] L. Salwinski, "The database of interacting proteins: 2004 update," *Nucleic Acids Res.*, vol. 32, no. 90001, pp. 449D–451, Jan. 2004.

[32] M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima, and M. Tanabe, "New approach for understanding genome variations in KEGG," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D590–D595, Jan. 2019.

[33] A. B. Dennis, C. Mark, C. Karen, K. M. Ilene, J. L. David, O. James, and W. S. Eric, "GenBank," *Nucleic Acids Res.*, vol. 41, pp. D36–D42, Nov. 2012.

[34] R. K. R. Kalathur, J. P. Pinto, M. A. Hernández-Prieto, R. S. R. Machado, D. Almeida, G. Chaurasia, and M. E. Futschik, "UniHI 7: An enhanced database for retrieval and interactive analysis of human molecular interaction networks," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D408–D414, Jan. 2014.

[35] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, A. Pon, K. Banco, C. Mak, V. Neveu, Y. Djoumbou, R. Eisner, A. C. Guo, and D. S. Wishart, "DrugBank 3.0: A comprehensive resource for 'Omics' research on drugs," *Nucleic Acids Res.*, vol. 39, pp. D1035–D1041, Jan. 2011.

[36] M. E. Fahey, M. J. Bennett, C. Mahon, S. Jager, L. Pache, D. Kumar, A. Shapiro, K. Rao, S. K. Chanda, C. S. Craik, A. D. Frankel, and N. J. Krogan, "GPS-Prot: A Web-based visualization platform for integrating host-pathogen interaction data," *BMC Bioinf.*, vol. 12, no. 298, pp. 1–13, 2011.

[37] H. W. Mewes, A. Ruepp, F. Theis, T. Rattei, M. Walter, D. Frishman, K. Suhre, M. Spannagl, K. F. X. Mayer, V. Stümpflen, and A. Antonov, "MIPS: Curated databases and comprehensive secondary data resources in 2010," *Nucleic Acids Res.*, vol. 39, pp. D220–D224, Jan. 2011.

[38] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms," *Protein Sci.*, vol. 28, no. 11, pp. 1947–1951, Nov. 2019.

[39] M. C. Costanzo, M. E. Crawford, J. E. Hirschman, J. E. Kranz, P. Olsen, L. S. Robertson, M. S. Skrzypek, B. R. Braun, K. L. Hopkins, P. Kondu, C. Lengieza, J. E. Lew-Smith, M. Tillberg, and J. I. Garrels, "YPDTM, PombePDTM, and WormPDTM: Model organism volumes of the Bio Knowledge TM library, an integrated resource for protein information," *Nucleic Acids Res.*, vol. 29, no. 1, pp. 75–79, 2001.

[40] W. E. Payne and J. I. Garrels, "Yeast protein database (YPD): A database for the complete proteome of saccharomyces cerevisiae," *Nucleic Acids Res.*, vol. 25, no. 1, pp. 57–62, Jan. 1997.

[41] R. Mosca, A. Céol, A. Stein, R. Olivella, and P. Aloy, "3did: A catalog of domain-based interactions of known three-dimensional structure," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D374–D379, Jan. 2014.

[42] A. Stein and P. Aloy, "Novel peptide-mediated interactions derived from high-resolution 3-dimensional structures," *PLOS Comput. Biol.*, vol. 6, no. 5, pp. 1–16, 2010.

[43] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. L. Sonnhammer, J. Tate, and M. Punta, "Pfam: The protein families database," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D222–D230, Jan. 2014.

[44] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. V. Mering, "STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, Jan. 2019.

[45] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, F. Zhang, S. Dolma, A. Willems, J. Coulombe-Huntington, A. Chatr-aryamontri, K. Dolinski, and M. Tyers, "The BioGRID interaction database: 2019 update," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D529–D541, Jan. 2019.

[46] A. Chatr-aryamontri, R. Oughtred, L. Boucher, J. Rust, C. Chang, N. K. Kolas, L. O'Donnell, S. Oster, C. Theesfeld, A. Sellam, C. Stark, B.-J. Breitkreutz, K. Dolinski, and M. Tyers, "The BioGRID interaction database: 2017 update," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D369–D379, Jan. 2017.

[47] H. Kumar, S. Srivastava, and P. Varadwaj, "Determination of protein-protein interaction through artificial neural network and support vector machine: Acomparative study," *Int. J. Comput. Biol.*, vol. 3, no. 2, pp. 37–43, 2014.

[48] A. Chinnasamy, A. Mittal, and W.-K. Sung, "Probabilistic prediction of protein–protein interactions from the protein sequences," *Comput. Biol. Med.*, vol. 36, no. 10, pp. 1143–1154, Oct. 2006.

[49] X. Li, L. Wang, and E. Sung, "AdaBoost with SVM-based component classifiers," *Eng. Appl. Artif. Intell.*, vol. 21, no. 5, pp. 785–795, Aug. 2008.

[50] L. Nanni, "Fusion of classifiers for predicting protein–protein interactions," *Neurocomputing*, vol. 68, pp. 289–296, Oct. 2005.

[51] Y. Guo, L. Yu, Z. Wen, and M. Li, "Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences," *Nucleic Acids Res.*, vol. 36, no. 9, pp. 3025–3030, May 2008.

[52] C. H. Q. Ding and I. Dubchak, "Multi-class protein fold recognition using support vector machines and neural networks," *Bioinformatics*, vol. 17, no. 4, pp. 349–358, Apr. 2001.

[53] W. Li, L. Jaroszewski, and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," *Bioinformatics*, vol. 17, no. 3, pp. 282–283, Mar. 2001.

[54] I. Xenarios, "DIP, the database of interacting proteins: A research tool for studying cellular networks of protein interactions," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 303–305, Jan. 2002.

[55] C. M. Deane, Ł. Salwiński, I. Xenarios, and D. Eisenberg, "Protein interactions: Two methods for assessment of the reliability of high throughput observations," *Mol. Cellular Proteomics*, vol. 1, no. 5, pp. 349–356, May 2002.

[56] W. Li and A. Godzik, "Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences," *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, Jul. 2006.

[57] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein-protein interactions based only on sequences information," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 11, pp. 4337–4341, Mar. 2007.

[58] E. Gasteiger, "ExPASy: The proteomics server for in-depth protein knowledge and analysis," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3784–3788, Jul. 2003.

[59] E. Coward, "Shufflet: Shuffling sequences while conserving the k-let counts," *Bioinformatics*, vol. 15, no. 12, pp. 1058–1059, Dec. 1999.

[60] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–26, 2011.

[61] Z. Wen, M. Li, Y. Li, Y. Guo, and K. Wang, "Delaunay triangulation with partial least squares projection to latent structures: A model for G-protein coupled receptors classification and fast structure recognition," *Amino Acids*, vol. 32, no. 2, pp. 277–283, Feb. 2007.

[62] K.-C. Chou and C.-T. Zhang, "Prediction of protein structural classes," *Crit. Rev. Biochemistry Mol. Biol.*, vol. 30, no. 4, pp. 275–349, 1995.

[63] A. Ben-Hur and W. S. Noble, "Choosing negative examples for the prediction of protein-protein interactions," *BMC Bioinf.*, vol. 7, no. S2, pp. 1–6, 2006.

[64] M.-G. Shi, J.-F. Xia, X.-L. Li, and D.-S. Huang, "Predicting protein–protein interactions from sequence using correlation coefficient and high-quality interaction dataset," *Amino Acids*, vol. 38, no. 3, pp. 891–899, Mar. 2010.

[65] X. Wu, "Prediction of yeast protein-protein interaction network: Insights from the gene ontology and annotations," *Nucleic Acids Res.*, vol. 34, no. 7, pp. 2137–2150, Apr. 2006.

[66] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[67] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[68] J. C. Rain, L. Selig, H. D. Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schächter, Y. Chemama, A. Labigne, and P. Legrain, "The protein-protein interaction map of Helicobacter pylori," *Nature*, vol. 409, no. 6817, pp. 211–215, 2001.

[69] J.-F. Xia, K. Han, and D.-S. Huang, "Sequence-based prediction of protein-protein interactions by means of rotation forest and autocorrelation descriptor," *Protein Peptide Lett.*, vol. 17, no. 1, pp. 137–145, Jan. 2010.

[70] W. Ma, Y. Cao, W. Bao, B. Yang, and Y. Chen, "ACT-SVM: Prediction of protein-protein interactions based on support vector basis model," *Sci. Program.*, vol. 2020, pp. 1–8, Jul. 2020.

[71] P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, and D. Plewczynski, "PPI_SVM: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables," *Cellular Mol. Biol. Lett.*, vol. 16, no. 2, pp. 264–278, Jan. 2011.

[72] J. Wojcik and V. Schachter, "Protein-protein interaction map inference using interacting domain profile pairs," *Bioinformatics*, vol. 17, pp. S296–S305, Jun. 2001.

[73] A. H. Land and A. G. Doig, "An automatic method of solving discrete programming problems," *Econometrica*, vol. 28, no. 3, pp. 497–520, Jul. 1960.

[74] T. Joachims, "Making large-scale SVM learning practical," MIT Press, Cambridge, MA, USA, Tech. Rep. 1998,28, 1998. [Online]. Available: https://www.cs.cornell.edu/people/tj/publications/joachims_99a.pdf

[75] X.-W. Chen and M. Liu, "Domain-based predictive models for protein-protein interaction prediction," *EURASIP J. Adv. Signal Process.*, vol. 2006, no. 1, pp. 1–8, Dec. 2006.

[76] D.-S. Han, "PreSPI: A domain combination based prediction system for protein-protein interaction," *Nucleic Acids Res.*, vol. 32, no. 21, pp. 6312–6320, Nov. 2004.

[77] H. Alashwal, S. Deris, and R. M. Othman, "A Bayesian kernel for the prediction of protein-protein interactions," *World Acad. Sci., Eng. Technol.*, vol. 3, no. 3, pp. 705–710, 2009.

[78] N. Zaki, "Prediction of protein-protein interactions using pairwise alignment and inter-domain linker region.," *Adv. Electr. Eng. Comput. Sci.*, vol. 39, no. 2008, pp. 635–645, 2008.

[79] W. K. Kim, J. Park, and J. K. Suh, "Large scale statistical prediction of protein-protein interaction by potentially interacting domain (PID) pair," *Genome Inf.*, vol. 13, pp. 42–50, 2002.

[80] S. Romero-Molina, Y. B. Ruiz-Blanco, M. Harms, J. Münch, and E. Sanchez-Garcia, "PPI-detect: A support vector machine model for sequence-based prediction of protein-protein interactions," *J. Comput. Chem.*, vol. 40, no. 11, pp. 1233–1242, Apr. 2019.

[81] Y.-R. Zou, A. H. Kottmann, M. Kuroda, I. Taniuchi, and D. R. Littman, "Function of the chemokine receptor CXCR4 in haematopoiesis and in cerebellar development," *Nature*, vol. 393, no. 6685, pp. 595–599, Jun. 1998.

[82] O. Zirafi, K. A. Kim, L. Ständker, K. B. Mohr, D. Sauter, A. Heigele, S. F. Kluge, E. Wiercinska, D. Chudziak, R. Richter, and B. Moepps, "Discovery and characterization of an endogenous CXCR4 antagonist," *Cell Rep.*, vol. 11, no. 5, pp. 737–747, 2015.

[83] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, Mar. 2011.

[84] R. D. Finn, B. L. Miller, J. Clements, and A. Bateman, "IPfam: A database of protein family and domain interactions found in the protein data bank," *Nucleic Acids Res.*, vol. 42, no. D1, pp. D364–D373, Jan. 2014.

[85] Y. B. Ruiz-Blanco, W. Paz, J. Green, and Y. Marrero-Ponce, "ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins," *BMC Bioinf.*, vol. 16, no. 1, pp. 1–15, Dec. 2015.

[86] L. H. Hall and L. B. Kier, "Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information," *J. Chem. Inf. Model.*, vol. 35, no. 6, pp. 1039–1045, Nov. 1995.

[87] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explor. Newslett.*, vol. 11, no. 1, pp. 10–18, 2009.

[88] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston, MA, USA: Addison-Wesley, 1989.

[89] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Dept. Comput. Sci. Inf. Eng., Univ. National Taiwan, Taipei, Taiwan, Tech. Rep., 2003.

[90] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, X. Luo, and A. Golshani, "PIPE: A protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs," *BMC Bioinf.*, vol. 7, no. 365, pp. 1–15, 2006.

[91] X. Liu, B. Liu, Z. Huang, T. Shi, Y. Chen, and J. Zhang, "SPPS: A sequence-based method for predicting probability of protein-protein interaction partners," *PLoS ONE*, vol. 7, no. 1, pp. 1–6, 2012.

[92] J. Y. Lo, J. A. Baker, P. J. Kornguth, and C. E. Floyd, "Computer-aided diagnosis of breast cancer: Artificial neural network approach for optimized merging of mammographic features," *Academic Radiol.*, vol. 2, no. 10, pp. 841–850, Oct. 1995.

[93] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digit. Signal Process.*, vol. 17, no. 4, pp. 694–701, Jul. 2007.

[94] R. Enyinnaya, "Predicting cancer-related proteins in protein-protein interaction networks using network approach and SMO-SVM algorithm," *Int. J. Comput. Appl.*, vol. 115, no. 3, pp. 5–9, Apr. 2015.

[95] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network based classification of breast cancer metastasis," *Mol. Syst. Biol.*, vol. 3, no. 140, pp. 1–10, 2007.

[96] I. H. I. Ahmed, "Computational prediction of host-pathogen protein-protein interactions," Ph.D. dissertation, South African Nat. Bioinf. Inst., Univ. Western Cape, Cape Town, Republic of South Africa, 2017.

[97] L. Dey, S. Chakraborty, and A. Mukhopadhyay, "Machine learning techniques for sequence-based prediction of viralehost interactions between SARS-CoV-2 and human proteins," *Biomed. J.*, vol. 43, no. 5, pp. 438–450, 2020.

ARIJIT CHAKRABORTY is currently pursuing the Ph.D. degree in computer science and engineering with the Maulana Abul Kalam Azad University of Technology, India. He is also an Assistant Professor with the Department of Computer Application, The Heritage Academy, Kolkata, India. His research interests include bioinformatics and machine intelligence.

SAJAL MITRA is currently associated with the Department of Computer Science and Engineering, Heritage Institute of Technology, Kolkata, India. He is also an Active Researcher in the field of machine learning, grid computing, and bioinformatics.

DEBASHIS DE (Senior Member, IEEE) received the M.Tech. degree in radio physics and electronics in 2002, and the Ph.D. degree in engineering from Jadavpur University, in 2005. He was a Research and Development Engineer with Telektronics. He is currently working as a Professor with the Department of Computer Science and Engineering, Moulana Abul Kalam Azad University of Technology, India, and an Adjunct Research Fellow with The University of Western Australia. His research interests include machine learning, low power nano device design for mobile application, and disaster management. He received the Boyscast Fellowship from the Department of Science and Technology, Government of India, to work with the Herriot-Watt University, U.K., and the Endeavour Fellowship Award from DEST, Australia, from 2008 to 2009 to work with The University of Western Australia. He was a recipient of the Young Scientist Award at New Delhi in 2005, and Istanbul in 2011, from the International Union of Radio Science, Belgium.

ANINDYA JYOTI PAL received the Ph.D. degree in engineering from Kalyani University, India. He has a teaching and research experience of 20 years. He worked as a Professor with the Department of Information Technology, Heritage Institute of Technology, Kolkata, India. He is currently associated with The University of Burdwan as a Controller of Examination. His research interests include soft computing and algorithm.

FERIAL GHAEMI received the Ph.D. degree in nanotechnology from the Institute for Advanced Technology, University of Putra, Malaysia, in 2015. She joined the Institute of Tropical Forestry and Forest Products in 2016, as a Postdoctoral Research Fellow. She is currently a Fellow Researcher with the Department of Chemical and Process Engineering, Faculty of Engineering and Built Environment, The National University of Malaysia, Malaysia. She has published many research articles in the prestigious CIJ journal. Her research interests include synthesizing various types of nanomaterials and their applications in drug discovery, polymer composites, and microextraction techniques.

ALI AHMADIAN (Member, IEEE) received the Ph.D. degree (Hons.) from University Putra Malaysia (UPM), in 2014. He is currently a Fellow Researcher with the Institute of Industry Revolution 4.0, UKM. As a Young Researcher, he is dedicated to research in applied mathematics. He worked on project related to drug delivery systems, acid hydrolysis in palm oil frond, carbon nanotubes dynamics, and Bloch equations and viscosity. He has successfully received 13 national and international research grants and selected as the 1% top reviewer in mathematics and compt1ter sciences recognized by Publons from 2017 to 2019. He is the author of more than 70 research articles published in reputed journals, including the IEEE TRANSACTION ON FUZZY SYSTEMS, *Fuzzy Sets and Systems*, *Communications in Nonlinear Sciences and Numerical Simulation*, the *Journal of Computational Physics*, and so on. He has presented his research works in 38 international conferences held in Canada, Serbia, China, Turkey, Malaysia, and UAE. His research interests include development of computational methods and models of problems arising in AI, biology, physics, and engineering under fuzzy and fractional calculus (FC). He was a member of programme committee in a number or international conferences in fuzzy field at Japan, China, Turkey, South Korea, and Malaysia. He is also a member of editorial board in *Progress in Fractional Differentiation and Applications* (Natural Science of Publishing). He serves as a Guest Editor for *Mathematical Methods in Applied Sciences*, *Advances in Mechanical Engineering* (SAGE), *Symmetry* (MDPl), *Frontier in Physics* (Frontiers), and the *International Journal of Hybrid Intelligence* (Inderscience Publishers). He also serves as a Referee for more than 70 reputed international journals.

MASSIMILIANO FERRARA received the master's degree *(cum laude)* in economics from the University of Messina, the master's degree *(cum laude)* from the University of Naples Federico II, the master's degree *(cum laude)* from the Scuola Normale Superiore di Pisa, and the Ph.D. degree *(cum laude)* from the University of Messina. He has been a Research Affiliate with the ICRIOS–Invernizzi Center for Research on Innovation, Organization, Strategy and Entrepreneurship, University Bocconi of Milan, since 2013, the President of Scientific Committee of the MEDAlics Research Center, and the Scientific Director of the DECISIONS Lab. He was a General Counsel of the FondazioneBanco di Napoli, the Vice Rector with the University for foreigners ''Dante Alighieri'' of Reggio Calabria, and the Head of Regione Calabria Department for Cultura, Research and Education. He was a Visiting Professor with Harvard University, Cambridge, MA, USA, Western Michigan University, USA, Morgan State University, Baltimora, MD, USA, the Northeastern University di Boston, USA, and recently with the Center for Dynamics of Dresden University of Technology, Germany. He is currently a Full Professor of mathematical economics, statistics, business analytics & decision theory, applied economics with the Mediterranea University of Reggio Calabria, where he is also the Chairman of the Department of Law, Economics & Human Sciences, and a Member of the Academic Senate. He is the author and coauthor of up 200 articles on peer-review and ISI journal and ten research monographs. He has been Knight Order of Merit of the Italian Republic since 2010 ''for international scientific merits''. He is in the prestigious US Encyclopedia Hmolpedia on thermodynamics and the theoretical and applied physics, for offering a decisive contribution to the creation and development of the scientific theory called ''Economic geometric dynamics''. He is also an Editor, a Co-Editor, an Associate Editor, and a Referee of reputable international scientific journal in economics, pure and applied mathematics.