ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 19(6): 749-753 (2023)

BIOINFORMATION
Discovery at the interface of physical and biological sciences

OPEN ACCESS GOLD

BIOMEDICAL INFORMATICS

www.bioinformation.net
Volume 19(6)

Research Article

Received June 1, 2023; Revised June 30, 2023; Accepted June 30, 2023, Published June 30, 2023

DOI: 10.6026/97320630019749

Declaration on Publication Ethics:
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Declaration on official E-mail:
The corresponding author declares that lifetime official e-mail from their institution is not available for all authors

Comments from readers:
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

Edited by P Kangueane
Citation: Kiran Kumar *et al*. Bioinformation 19(6): 749-753 (2023)

Prediction of transient and permanent protein interactions using AI methods

A. Kiran Kumar*, Syed Mohammad Shayez Karim, Mayank Kumar & Ravindranath Singh Rathore*

Department of Bioinformatics, Central University of South Bihar, Gaya, Bihar-824236, India; *Corresponding authors

Institution URL:
https://www.cusb.ac.in

Author Contacts:
Kiran Kumar - E-mail: kirankumar@cusb.ac.in
Shayez Karim - E-mail: shayazkarim@cusb.ac.in
Mayank Kumar - E-mail: mayank@cusb.ac.in
R.S. Rathore - E-mail: rsrathore@cusb.ac.in

Abstract:
Protein-protein interactions (PPIs) can be classified as permanent or transient interactions based on their stability or lifetime. Understanding the precise details of such protein interactions will pave the way for the discovery of inhibitors and for understanding the nature and function of PPIs. In the present work, 43 relevant physicochemical, geometrical and structural features were calculated for a curated dataset from the literature, comprising of 402 protein-protein complexes of permanent and transient categories, and 5 different

Supervised Machine Learning models were developed with *Scikit-learn* to predict transient and permanent PPI. Additionally, deep learning method with Artificial Neural Network was also performed using *Tensor Flow* and *Keras*. Predicted models achieved accuracy ranging from 76.54% to 82.71% and k-NN has achieved the highest accuracy. Detailed analysis of these methods revealed that Interface areas such as Percent interface accessible area, Interface accessible area and Total interface area and the parameters defining the shape of the PPI interface such as Planarity, Eccentricity and Circularity are the most discriminating factors between these two categories. The present method could serve as an effective tool to understand the mechanism of protein association and to predict the transient and permanent interactions, which could supplement the costly and time-consuming experimental techniques.

Key words: Transient and Permanent Protein-Protein Interactions; Machine Learning; *Scikit-learn*; Deep Learning; *Tensor Flow*.

Background:

A host of biological and cellular activities, such as gene replication, transcription, translation, cell cycle regulation, signal transmission, and immune response, rely on protein-protein interactions. Protein-protein interactions (PPIs) are vital for understanding how proteins work together in the cell to accomplish biological tasks in a coordinated manner **[1, 2]**. An estimated 130,000 to 650,000 different types of protein–protein interactions exist in human cells **[3-5]**. Such interactions belong to permanent or transient categories of interactions, which play a specific role in cellular activities **[6, 7]**. Permanent complexes such as enzyme-inhibitor, antigen-antibody, and oligomeric enzyme are composed of proteins that bind tightly and permanently, whereas transient complexes weakly associate and form just temporarily to produce specific effects like signal transduction, disease related pathways and cell cycle **[8, 9]**. These interactions are distinguished by their dissociation constant (Kd) as permanent complexes having dissociation value in the nM range (1×10^{-9} M) or lower **[10, 11]**, whereas transient complexes have dissociation constant in the µM range or higher (1×10^{-6} M) **[12-14]**. The ability to manipulate these protein–protein interactions could be useful in the development of PPI modulators, which could open up new avenues for biologics research **[15, 16]**. A deep structural understanding of such complexes at the atomic level will enhance our knowledge of biological processes and may facilitate biomedical and biotechnological interventions easier. Earlier, investigations have been carried out primarily using sequence-based features **[17-20]** to elucidate the differences between permanent and transient protein interactions. Permanent interaction sites have been found to possess more hydrophobic residues, more conserved, and their interfaces contain fewer gaps in multiple sequence alignments of protein families. On the other hand, transient interfaces have more polar residues, and they form smaller interfaces than permanent interfaces **[19]**. Machine-learning techniques have proven to be effective in predicting and distinguishing different types of PPIs **[21-23]**. Recently, a wide number of state-of-the-art techniques to predict protein-protein interactions have been reviewed **[24]**. In the present study, we have employed several supervised machine learning and deep learning methods to classify transient and permanent interactions by calculating various physicochemical, geometrical and structural factors that define transient and permanent protein interactions. In our calculations, different properties like Percent interface accessible area, Interface accessible area, and Total interface area, Planarity, Circularity and Eccentricity were discovered to be capable of discriminating between transient and permanent protein interactions. Our approaches of diverse supervised machine learning algorithms and Artificial Neural Networks (ANN) were able to differentiate 402 protein–protein complexes with an accuracy of 76.54 to 82.71%.

Materials and methods:

Dataset preparation and processing:

Dataset of protein complexes to study transient and permanent interactions were compiled from the literature **[19-21]**. The dataset contains a total of 402 transient and permanent protein complexes containing 201 complexes belonging to each category (List of PDB entries included in Supplementary Table S1).Various categories of structural, physicochemical and geometrical descriptors were calculated using 2P2I inspector **[25]**. We have calculated a total of 43 different features such as total interface area, gap volume, percent interface accessible surface area, neutral/polar/nonpolar contribution, planarity, circularity, eccentricity and others (listed in **Figure 1**). Missing data and outliers were cleaned and data were pre-processed using *Scikit-learn* Standard Scaler utility. All descriptors were rescaled between 0 and 1.

Supervised Machine Learning with Python:

Scikit-learn was used to construct the classification models and training the data to determine the best parameters for the training model using different algorithms such as k-Nearest Neighbour (k-NN) **[26]**, Logistic Regression [27], Decision Tree **[28]**, Random Forest **[29]** and Support Vector Machine (SVM) **[30]**. *Pandas v1.1.5*, *matplotlib v3.2.2*, *NumPy v1.19.5*, *SciPy v1.4.1*, *Scikit-learn v0.22.2* **[31]**, and *seaborn v0.11.2* were used to perform the machine learning. In all our models, the datasets were divided into training and test sets, in the ratio of 80:20. In k-NN, several distance metrics were evaluated in *Scikit-learn*, including k = 1 to 5 nearest neighbours, to predict the data. In Random Forest, the number of decision trees was set as 500. For Logistic Regression, different logistic regression classifiers have been employed by varying C value from 100 to 1000 and the best accuracy was achieved with C=500. The precision score, sensitivity or recall and F1 score, which is the weighted average of both the precision score and recall were calculated for each algorithm (detail description about these parameters provided in the supplementary material). These performance measurements were calculated for each class that is transient and permanent and the geometric mean (G-mean) of sensitivity and specificity was also computed (**Table 1**). We performed variable importance calculation using Boruta and Random Forest in Python as shown in the **(Figure 1).**

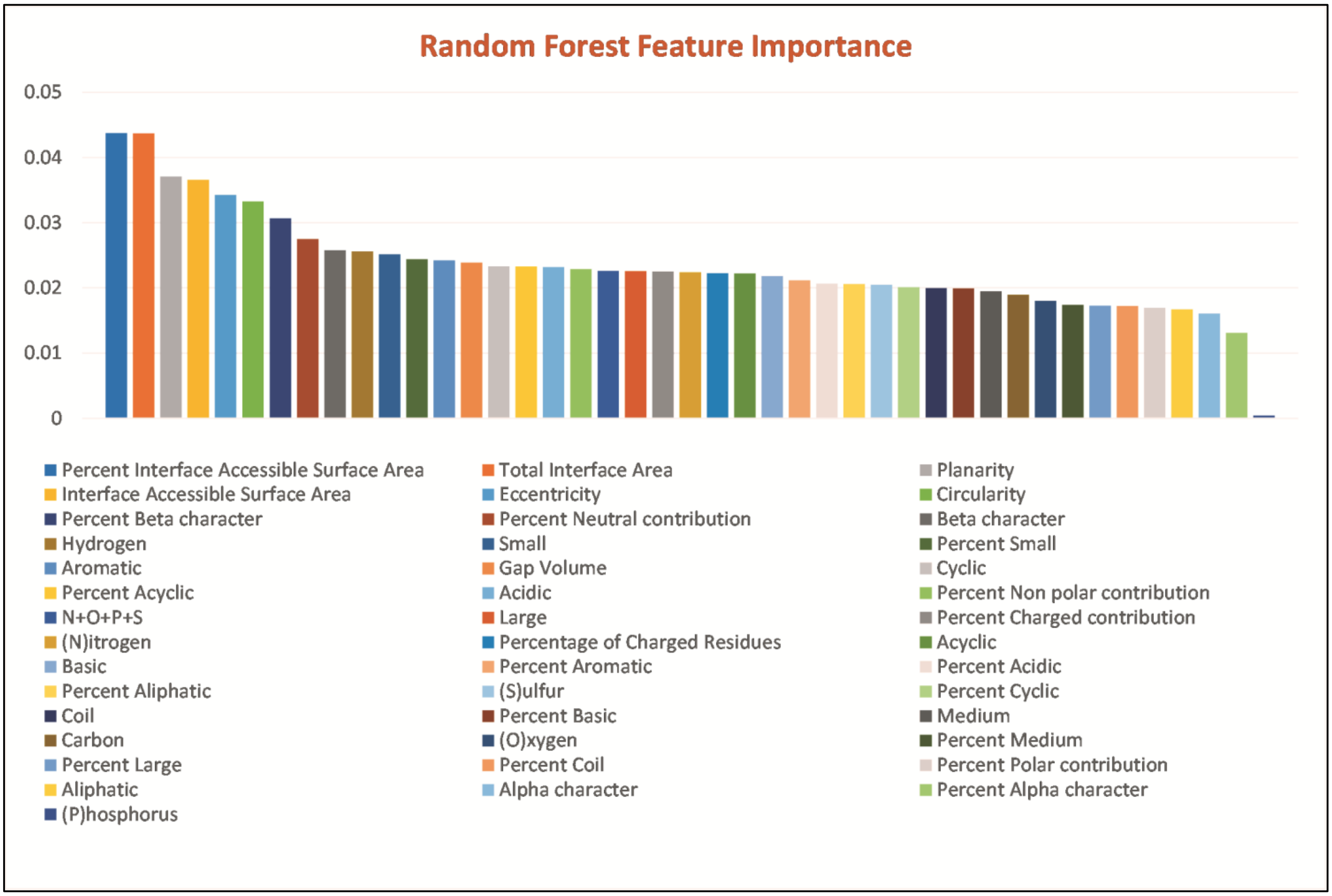

Figure 1: Feature importance plot performed with Random Forest. Feature significance score is displayed on Y-axis. Definition of features is as per reference **[25]**

Table 1: Performance measurements of Machine learning models obtained With Scikit-learn.

ML Method	Accuracy	Class	Sensitivity	G-Mean of Sensitivity and Specificity	Precision	F1 Score
k-NN		Transient	0.804	0.804	0.846	0.824
		Permanent	0.85	0.804	0.809	0.828
	82.71	Average	0.827	0.804	0.827	0.826
Random Forest		Transient	0.756	0.813	0.861	0.805
		Permanent	0.875	0.813	0.778	0.823
	81.48	Average	0.815	0.813	0.819	0.814
Logistic Regression		Transient	0.804	0.801	0.804	0.804
		Permanent	0.8	0.801	0.8	0.8
	80.24	Average	0.802	0.801	0.802	0.802
Decision Tree		Transient	0.75	0.776	0.789	0.769
		Permanent	0.804	0.776	0.767	0.785
	77.77	Average	0.777	0.776	0.778	0.779
SVM		Transient	0.744	0.766	0.8	0.77
	76.54	Permanent	0.789	0.766	0.731	0.758
		Average	0.766	0.766	0.765	0.764

Deep Learning with Tensor Flow:

We used *Tensor Flow* and *Keras* to implement the deep learning. Deep learning models **[32]** are made up of multiple computational layers that process the input in a hierarchical manner. Each layer takes an input and outputs a non-linear function of a weighted linear combination of the input values. A deep architecture is created when the output of one processing layer becomes an input to the next processing layer. Networks with two hidden layers were adopted to compare their performance in our study. We used ReLU as an activation function for the two hidden layers and sigmoid function for the output layer. As earlier, the data were divided into training and test set in 80:20 ratios.

Results and discussion:

Based on 43 descriptors, several machine learning and deep learning methods were attempted to arrive at consensus results. The accuracy of the methods and other performance evaluation metrics were calculated and reported in Table 1. The accuracy of different methods achieved, range between 76.54% to 82.71% prediction of the data using physicochemical, geometrical and structural features. The highest accuracy of 82.71% was achieved with k-NN (**Table 1**). The values of precision and F1 score of the method were 0.827 and 0.826, respectively. The other supervised machine learning algorithms – Random Forest, Logistic Regression, Decision Trees and SVM have yielded accuracies of 81.48%, 80.24%, 77.77% and 76.54%, respectively. The deep learning with ANN achieved the accuracy of 79% with 500 epochs and with *adam* as the optimizer for 43 input dimensions. To elucidate the relative feature importance in transient and permanent categories, the feature contributions were also calculated. One of the most discriminating category of features in this classification procedure is interface areas, namely, Percent interface accessible area, Interface accessible area and Total interface area with feature importance score of 0.0437, 0.0436 and 0.036, respectively. The value of these parameters for transient PPI have been observed significantly lower as compared to permanent PPI. The average value of Percent interface accessible area, Total interface area & Interface accessible area for transient PPI category have been observed to be 10.98%, 2594.4Å^2 & 1291.2Å^2, respectively, as against 15.11%, 3819.1Å^2 & 1911.5Å^2, respectively for in permanent PPI category.

The second most important category of discriminating features is the one that describe the shape of interface such as Planarity, Eccentricity and Circularity with feature importance scores of 0.037, 0.034 & 0.033, respectively. The Planarity describes the rough or bent interface **[25, 33]** and calculated as root mean square deviation (RMSD) for all interface atoms from the best fitted least square plane of all the interface atoms. The average planarity coefficient in transient PPI category varies between 0.29-7.2 Å (Avg. 3.02 Å) as compared to 0.57-10.6 Å (Avg. 3.8 Å) in permanent PPI category. Eccentricity (roundness of the interface and opposite to the curvature) suggest slightly low curvature in transient category, 0.2-0.99 (Avg. 0.73) than in permanent PPI, 0.12-0.979 (Avg. 0.68). Another such measure i.e., Circularity coefficient is also found to be slightly lower that varies between 0.123-0.98 (Avg. 0.61) in transient PPI than 0.20-0.99 (Avg. 0.68) in permanent PPI category. A related parameter of interface shape is the Gap volume with a feature importance score of 0.023. In transient categories the average gap volume was slightly higher 7775.2 Å^3 as compared to permanent category having a value of 7717.1 Å^3. The third most categories of discriminating features are the percentage of beta character with a feature important score of 0.031. For transient PPI its value varies in between 0-100 (Avg. 21.6), and in permanent PPI the value is higher, which has been found to be 0-94 (Avg. 26.6).

Conclusion:

Transient and permanent protein–protein interactions are significant in many biological processes. In the present work, we used a dataset, compiled from the literature and extracted physicochemical, geometrical and structural features from each of the 201 permanent and transient protein-protein complexes. Interface areas, shape of the interface and percent beta character are the three distinct categories of features, which prominently discriminate transient and permanent interactions. The method we proposed here could be useful in engineering permanent or transient PPIs, notably in the conversion of permanent docking interfaces to transient docking interfaces or vice versa using interface mutations **[16]**. The ability to manipulate these protein–protein interactions should aid in structure-aided biologics discovery. In addition, the present methodology may also be used to classify other similar types of interactions such as protein-DNA and protein-RNA interactions.

Associated Data:

Supplementary Materials

References:

[1] Sudha G *et al. Prog Biophys Mol Biol*. 2014 **116**:141. [PMID: 25077409]

[2] Nicod C A *et al. Curr Opin Microbiol.* 2017 **39**:7. [PMID: 28806587]

[3] Venkatesan K *et al. Nat Methods*. 2009 **6**:83. [PMID: 19060904]

[4] Nero T.L *et al. Nat Rev Cancer*. 2014 **14**:248. [PMID: 24622521]

[5] Stumpf M.P *et al. Proc Natl Acad Sci U S A.* 2008 **105**:6959. [PMID: 18474861]

[6] Ngounou Wetie *et al. Proteomics* 2013 **13**:538. [PMID: 23193082]

[7] Keskin O *et al. Chem Rev*. 2016 **116**:4884. [PMID: 27074302]

[8] Kastritis P.L *et al. Protein Sci* 2011 **20**:482**.** [PMID: 21213247]

[9] Perkins, J.R *et al. Structure*. 2010 **18**:1233. [PMID: 20947012

[10] Ding, Z. and D. Kihara, *Sci Rep*. 2019 **9**:8740. [PMID: 31217453]

[11] Jayashree, S *et al. Biol Direct.* 2019 **14**:1. [PMID: 30646935]

[12] Ji, L. *et al. Energies.* 2016 **9**:898 [https://www.mdpi.com/1996-1073/9/11/898]

[13] Kim, D.H *et al. PLoS Biol*. 2018 **16**:e2006660. [PMID: 30543635]

[14] Kathera, C *et al. Oncotarget*. 2017 **8**:27593. [PMID: 28187440

[15] Villoutreix, B.O *et al. Mol Inform*. 2014 **33**:414. [PMID: 25254076]

[16] Plattner, N *et al. Nat Chem*. 2017 **9**:1005. [PMID: 28937668]

[17] Markmiller, S *et al. Cell*. 2018 **172**:590. [PMID: 29373831]

[18] Peng, X., *et al. Brief Bioinform*. 2017 **18**:798. [PMID: 27444371]

[19] La, D *et al. Proteins.* 2013 **81**:805. [PMID: 23239312]

[20] Block, P *et al. Proteins*. 2006 **65**:607. [PMID: 16955490]

[21] Yugandhar, K. & M.M. Gromiha, *Proteins*. 2014 **82**:2088. [PMID: 24648146]

[22] Rong, Y *et al. Analyst*. 2018 **143**:2066. [PMID: 29629449]

[23] Paul George, A.A *et al. BMC Bioinformatics*. 2020 **21**:124. [PMID: 32216745]

[24] Ding, Z. & D. Kihara, *Curr Protoc Protein Sci,* 2018 **93**:e62. [PMID: 29927082]

[25] Basse, M.J *et al. Database (Oxford).* 2016 **41**:D824. [PMID: 26980515]

[26] Zhang, S *et al. IEEE Transactions on Neural Networks and Learning Systems.* 2018 **29**:1774. [PMID: 28422666]

[27] Wang, Q. Q *et al. Zhonghua yu fang yi xue za zhi (Chinese Journal of Preventive Medicine).* 2019 **53**:955. [PMID: 31474082]

[28] Quinlan, J.R, *Machine Learning*. 2004 **1**:81. [https://doi.org/10.1023/A:1022643204877]

[29] Breiman, L, *Machine Learning*, 2001 **45**:5. [https://doi.org/10.1023/A:1010933404324]

[30] Pradhan, Ashis. *IJETAE.* 2012 **2**:82.

[31] Pedregosa, F *et al. J. Mach. Learn. Res*. 2011 **12**:2825

[32] LeCun, Y *et al. Nature*, 2015 **521**:436. [PMID: 26017442]

[33] Nooren, I.M. & J.M. Thornton, *J Mol Biol*. 2003 **325**:991. [PMID: 12527304]

Supplementary materials:

Table S: Dataset (PDB Ids) of 402 transient and permanent protein-protein complexes.

Transient Protein Interaction Dataset								
1a00	1cjd	1h2t	1o0v	2ckl	2pmw	3gb8	4drx	5hzp
1a0o	1cmi	1h4r	1oan	2d07	2pmz	3hd7	4fqx	5jne
1a2w	1cqp	1hl6	1occ	2dd8	2qkl	3hu1	4g8f	5k93
1a37	1csg	1HQM	1rkc	2egd	2ql5	3i6l	4gdk	5kdm
1a7x	1d8t	1hvv	1S5L	2erj	2qxv	3jua	4hsu	5sy8
1a8M	1dev	1hxb	1sko	2ewy	2r83	3kwq	4ifd	5vok
1ab9	1dm4	1i3q	1svx	2fntA	2vhs	3kzi	4il6	5wsv
1afv	1e9h	1ifd	1tvp	2gg2	2wii	3m99	4jk1	6cnr
1agr	1eba	1ivo	1uwh	2gic	2z31B	3nc1	4k71	6ea7
1ahw	1egw	1iw7	1vf5	2gro	3a0b	3prx	4k94	
1an7	1ejl	1izl	1vgl	2h4m	3al4	3rk2	4m40	
1ao6	1eo8A	1izn	1w26	2hwn	3b8e	3s4s	4mng	
1aoi	1es7	1jh5	1wp8	2hxY	3bpo	3uzq	4qrs	
1aqd	1ezv	1jqj	1x79	2iae	3bw1	3vbf	4qyz	
1azz	1f66	1jt3	1xu7	2iff	3bwu	3vbfC	4tvp	
1b34	1fjg	1k8k	1z2c	2jjs	3csy	3w97	4w6bA	
1b3u	1foc	1kla	1z8j	2kwf	3d85	3wmm	4wxv	
1b50	1fs1	1l1o	1zru	2l2i	3ddc	3wod	4y6a	
1bcc	1g3j	1ldj	1zy8	2lr1	3dhg	3wxe	5ayw	
1be3	1g8q	1lm8	1zys	2mre	3dx9	3zk6	5c0z	
1bmf	1gag	1m4r	2a2y	2nl9A	3e4z	3zni	5dis	
1bqh	1gfw	1m63	2a73	2O8v	3e7a	4bsv	5dn6	
1c9b	1ggk	1mg2	2b4j	2oj5	3fwb	4c8q	5fv1	
1cfm	1gzh	1nys	2b5l	2pm6	3g7v	4cc9	5hlu	

Permanent Protein Interaction Dataset								
1A3C	1FCD	1MJL	3GRS	1OC0	1DCE	1JSG	2AHJ	1KXP
1A6D	1FIP	1MKA	3GTU	1OPH	1DJ7	1JV2	2CCY	1BVN
1A9X	1FM2	1MOQ	3PCG	1P2C	1E9Z	1KBA	2ILK	1DFJ
1AD3	1FRO	1NOX	3PGH	1PXV	1EFV	1KFU	1REQ	1DQJ
1AF5	1FS0	1NSY	3RUB	1R0R	1EG9	1KPF	1RFB	1EAW
1AFW	1FXW	1OAC	3SDH	1RV6	1EP3	1HXM	1RPO	1EER
1AHJ	1G72	1OPY	3SSI	1T6B	1EUD	1HZZ	1RTH	1EMV
1AJS	1G8J	1OTP	4KBP	1UUG	1BSR	1I1Q	1SES	1EZU
1ALK	1G8K	1PAU	4MON	1VFB	1BUO	1I3R	1SKY	2JEL
1AMK	1GK9	1PGT	5CSM	1WDW	1CCW	1I4F	1SLT	
1AOM	1GO3	1PHN	5TMP	1WEJ	1CD1	1I7B	2J0T	
1AOR	1GOT	1PRE	9WGA	1YVB	1CG2	1IAK	1F34	
1AQ6	1GVP	1PUC	1ACB	1ZLI	1CHM	2I9B	1FLE	
1AUI	1H2R	1QDU	1AHW	2ABZ	1CMB	1SMN	1FSK	
1AUO	1HCN	1QGW	1ATN	2B42	2I25	1SMT	1GPW	
1AW8	1HFE	1QH1	1AVX	2GOX	1ICW	1SOX	1GXD	
1B4U	1HGE	1QLA	1AY7	2HRK	1IHF	1SPP	1HCF	
1B5F	1HJR	1QOP	1BJ1	1CP2	1IMB	1TOX	1I2M	
1B7Y	1HLR	1QS0	1BRS	1CSH	1IRD	1TRK	1IBR	
1BAM	1HR6	1QTN	1M10	1CTT	1ISA	1TYS	1IQD	
1BIF	1HSS	1KXQ	1MAH	1CZJ	1ISO	1UBY	1JIW	
1BMV	2LTN	2RSP	1NB5	1D09	1JHG	1UTG	1JPS	
1KVD	1LUC	2TCT	1NCA	1D2V	1JK0	1WGJ	1JTG	
1EXB	1LYN	2TGI	1NSN	1DAA	1JRO	1XSO	1K5D	

Parameters for evaluation of performance of the machine learning methods

True Positive Rate (TPR)/Sensitivity/Hit Rate/Recall $= \frac{TP}{TP+FN}$ (1)

True Negative Rate (TNR)/Specificity/Selectivity $= \frac{TN}{FP + TN}$ (2)

FPR (False Positive Rate) $= \frac{FP}{TN + FP}$ (3)

FNR (False Negative Rate) $= \frac{FN}{FN+TP}$ (4)

Precision $= \frac{TP}{TP+FP}$ (5)

Recall $=$ TPR (6)

F1 Score $= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ (7)

The accuracy is defined as:

Acc $= \frac{TP+TN}{TP+TN+FP+FN}$ (8)

Where TP stands for true positives. TN for true negatives, FP for false positives, and FN for false negatives, predicted by the classifier. The F1 score is defined as the harmonic mean of precision and recall: