IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 11, 2022, accepted April 26, 2022, date of publication May 3, 2022, date of current version May 12, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3172278

Prediction of Aptamer Protein Interaction Using Random Forest Algorithm

**N. MANJU[1], C. M. SAMIHA[1], S. P. PAVAN KUMAR[2],
H. L. GURURAJ[2], (Senior Member, IEEE),
AND FRANCESCO FLAMMINI[3], (Senior Member, IEEE)**

[1]Department of Information Science and Engineering, JSS Science and Technology University, Mysore 570006, India
[2]Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysore 570002, India
[3]IDSIA USI-SUPSI, University of Applied Sciences and Arts of Southern Switzerland, 6928 Manno, Switzerland

Corresponding author: Francesco Flammini (francesco.flammini@supsi.ch)

ABSTRACT Aptamers are oligonucleotides that may attach to amino acids, polypeptide, tiny compounds, allergens and living cell membrane. Therapeutics, bio sensing and diagnostics are all sectors where the aptamers may be used. In this work, we present a model based on Random Forest Algorithms to predict the interaction of aptamer and target proteins by combining their most prominent characteristics. Amino Acid Composition and Psuedo Amino Acid Composition were utilized to express desired data by employing physicochemical and structural features of the amino acids. The dominant features were selected using feature importance classifiers such as random forest and eXtreme Gradient Boosting. Compared to these, principal component analysis techniques yielded a good feature set. As a result, 98% accuracy is obtained for 50 principal components. Many relevant characteristics in defining aptamer target protein interactions were discovered after analysing the best set of features. Our prediction approach is expected to become a valuable tool for discovering aptamer-target interactions, and the traits chosen and studied in this work might give helpful insight into the process of Aptamer Protein interactions.

INDEX TERMS Aptamers, random forest algorithm (RFA), random forest feature importance (RFFI), XGBoost feature importance (XGBFI), principal component analysis (PCA), psuedo amino acid composition (PseAAC).

I. INTRODUCTION

Aptamers are single-stranded DNA or RNA molecules that may attach to amino acids, polypeptide, tiny compounds, allergens and living cell membrane as mentioned in reference[1]. Due to their proclivity for forming side chains and solitary loops, aptamers can take on a variety of geometries. They're incredibly adaptable and have a high level of specificity and selectivity when it comes to binding targets. Aptamer affinity is deduced from its tertiary structure instead of its basic nucleotide. Receptor detection and binding are affected by dimensionality, texture affinities, water molecules etc.

Benefits of aptamers are: a) Aptamer synthesis: Aptamers may be produced in large quantities with high precision and repeatability using chemical processes once they've been chosen; b) Firmness: Upon renaturation, aptamers regain their natural structure and may attach to targets. They may be utilised in a variety of test circumstances; c) Low resistance: aptamers appear to be low-immunogenic and low-toxic substances because DNA molecules are not typically recognised as third parties by the body's immune system as mentioned in reference [1]; d) Diversification of spot: Aptamers can be produced in adequate numbers in the case when poisons or other compounds that do not provoke robust immune responses. It has a high affinity and specificity for ligands that antibodies can't identify such ions or tiny compounds.

Aptamers can be synthesised with the biological process. SELEX is a technique that selects compatible oligonucleotide for a particular target from a huge oligonucleotide pool. Bacteria samples will be added to the large aptamers pool and it will be washed with the chemical solution. Aptamers that do not bind to the desired target are eliminated, while those that do are enlarged for the next process as mentioned in reference [1]. By using polymerase chain reaction (PCR) method billions of copies of a DNA sample will be created in a short amount of time and then fluorescent will be added to this to identify the pathogen in food.

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

The pool of possible interactions would be reduced to a slice of possible typical pairings using computational approaches. Interaction pair of protein-protein and aptamer-protein can be predicted with the aid of machine learning. The result will act as a platform for additional laboratory research. Rapid innovations have indeed resulted in the creation of networks, including all of the protein-protein interactions which may be used to identify protein complexes in certain disorders using computational approaches as mentioned in reference [2].

SVM and KNN are a type of supervised ML technique that would be employed to solve classification and regression problems [26], [27]. It is employed to solve categorization difficulties. Researchers depict every piece of data as just a point in n-dimensional spaces in the SVM method, with feature vector becoming the values of a certain coordinates. Then we accomplish categorization by locating the hyperplane, which clearly distinguishes the 2 groups. Where as in KNN, initially number K of the neighbor will be selected and Euclidean distance will be calculated. As per the calculated distance, take the K-nearest neighbor and count the number of data points in the category. Finally, assign the new data points to that category for which the number of the neighbor is high. Extreme Gradient Boosting and Random forest algorithm are tree based structures which performs well on huge dataset. Therefore, with the help of various machine learning techniques, PPI and API can be predicted by considering Physiochemical properties of the proteins.

II. LITERATURE SURVEY

Analysis of the Physiochemical properties of the proteins is essential to predict the protein aptamer interaction. With the help of SELEX (Systematic Evolution of Ligands by Exponential Enrichment) cycle new aptamers can be identified but it is quite costly. By employing machine learning algorithms, protein and protein-protein interactions can be predicted. DNA aptamer is identified for inconsiderable lung carcinoma cells with plasma membrane markers as mentioned in reference [3]. A549 i.e., the adenocarcinoma cell line of human was utilized for selection with multiple cycles in SELEX whereas blood leukocytes are employed for negative selection and CD90 antibody cells employed for constructive selection. The acquired patterns were issued to in silico study, depending upon binding affinity, stability and structural motifs.

The deliberated phylogenetic tree exhibited that A549 cell and aptamer A155_18 have an immediate structural relationship with strong binding affinity. Therefore, the aptamer A155_18 is considered as one of the diagnostic tools, recognizing NSLC cells. By cell SELEX method HF3_58 and HA5_68 are the two aptamers which are generated in opposition to A2780T which is a paclitaxel resistant ovarian cancer cell line with robust affinity and high selectivity [4]. On the cell surface of A2780T, the two aptamers were explanatorily exhibited to be two different glycoproteins. The aptamers were bound independently of divalent ions, temperature and nuclease. Aptamers applied for the identification of drug-resistant ovarian cancer cells in human serum as mentioned in reference [2].

Angiogenesis takes an important part in the extension and expansion of cancer cells, but Ang2 has a unique role in regulating angiogenesis [5]. Computational simulation is carried out to show aptamers with high binding ability for Ang2 (angiopoietin-2) as mentioned in reference [5]. This is the initial study against Ang2 using in silico selection with the ZRANK scoring function, which helps to maximize the coherence of selected aptamers with good target-binding ability. Here obtained RNA sequences are converted into the 3 D structure and then the ZRANK and ZDOCK functions are applied. Based on ZRANK, the top three sequences are selected. 189 sequences were generated with two point mutations and it is simulated with Ang2. Later, in order to test three mutant sequences of maximum ZRANK scores SPR (Surface Plasmon Resonance) is used.

15-mer aptamers were traced for cytochrome P450 51A1 by employing simulation and molecular docking as mentioned in reference [6]. The approach employed here is in silico where three stages were involved, i.e., identifying a possible binding area, plotting the identification and systemic portion of the aptamers and estimating the experimental affinity. SPR biosensors showed that investigational estimation of the synthesized aptamers interaction with cytochrome P450 51A1 with Kd values in the range of $[10^{-}6 - 10^{-}7]$ M as mentioned in reference [6].

For the estrogen receptor (ER) alpha an aptamer is discovered by employing computational docking as mentioned in reference [7]. ERE has a power to form solid hairpins and it was considered as one of the main benchmark to acquire aptamer-alike sequences. Single stranded RNA analogs of human estrogen response elements (EREs) were formed and their possibility to appear as ER aptamer was inspected by employing HADDOCK, PatchDock and AutoDock Vina. The entire work is verified by calculating the thermodynamic constants of ER. A candidate RNA owns a binding constant (Ka) of $1.02 \pm 0.1 \times$ 108 M-1 is selected as an ER-aptamer based on the in silico and in vitro results. The high specificity and affinity can be used in identification of ER in breast cancer and associated diseases as mentioned in reference [7].

The discrete prognostic model is necessary to forecast PPIs of latest hosts and virus because the available computational method is restricted to single virus and host [8]. Most statistical approaches for forecasting PPIs are designed for interactions in species instead of interactions in the middle of species such as virus-host cell protein interactions. Despite poor sequence similarity between test and training dataset proteins, the forecasting model performed well, comparable to the best results of other approaches for single virus-host PPIs. The forecasting model is evaluated on a separate dataset of virus-host PPIs that were not included in the model's testing and have a poor sequence similarity to any protein in the model's training datasets.

More enhanced foregoing method of presentation and identified strong attribute for forecasting virus-human protein interactivity has been discussed in reference [9]. RFAT, AC and FDAT is represented in human proteins and virus in vector of the function. The most popular method to forecast PPIs is AC. However, in the middle of host and virus proteins FDAT and RFAT was employed in this research work to forecast interactions of virus-host protein. SVM model was created in order to forecast human proteins that interact with HPV and HCV and it yielded 66.9% accuracy. Using the new features and representation process, a large volume of data involved in PPIs is expressed in function vectors of minimal and fixed dimension, but still it achieved significantly better performance than previous computational methods. Other forms of heterogeneous PPIs can be predicted using the features and representation process.

A number of statistical approaches for locating aptamers have been proposed. Many of those approaches, though, can't be used to find latest aptamers for a target as long as they're either classifier for deciding whether a specific set of RNA sequences and protein interacts or they're restricted to a single target. The latest computational technique has been represented to construct strong RNA aptamers for a protein target employing many attributes of proteins and secondary structure of RNA has been discussed in reference [10]. A random forest model is built by choosing few features which showcased good performance in both independent testing and cross validation. The proposed method decreases the cost spent on in vitro and time by considerably decreasing the nucleic acid sequence pool's primary dimension.

AptaNet is unusual in that it predicts API by combining sequence-based functionality for oligonucleotides with conformational and physicochemical properties for goals has been discussed in reference [11]. AptaNet outperforms other approaches in terms of precision. AptaNet has shown the ability to offer biological observations into the existence of APIs, which can be beneficial to all oligonucleotides researchers and biologists. The forecasting model was built employing a DNN for 3404 instances, 640 features, and the random forest algorithm was used to pick the best attributes. As a result, the research dataset achieved a 91.38% accuracy. AptaNet outperformed the competition on a built-in aptamer-protein benchmark dataset.

By taking into account essential properties of protein molecules, the Random Forest model may be utilised to predict possible RNA aptamers for a protein target has been discussed in reference [12]. With the aid of RF model, ranks of 38,327 RNA sequences were identified in its secondary structure and dominant 10 RNA aptamers were selected. Later, the built SVM and RF prediction model is tested using cross validation and independent testing method. RF and SVM model achieved 97.76% and 96.08% accuracy for 10 fold respectively.

Ensemble classifier can be used to predict the aptamer protein interaction has been discussed in reference [13] because it gives good result by combining numerous basic classifiers. Feature extraction is carried out for 2900 protein aptamer interaction pair which is collected from an aptamer database with pseudo K-tuple nucleotide composition methods. The Relief-Incremental Feature Selection (IFS) was used to select prominent features. The ensemble model achieved 73.2% of accuracy after performing feature selection method.

The properties produced from the Pseudo Amino Acid Composition technique can be used to forecast aptamer-target interaction pairs has been discussed in reference [14]. The ideal 220 features were picked for 2900 instances employing the maximum relevance minimal redundancy (mRMR) approach and the incremental feature selection (IFS) approach, and the predictor was built using Random Forest. Achieved 81.34% and 77.41% accuracy for the training and testing dataset respectively.

Interactions between aptamers and proteins are significant in physiological activities and molecular identification. Despite various uses of aptamers, determining AP interaction pairs is difficult and restricted. With the help of sparse auto encoder the features of target sequences can be extracted has been discussed in reference [15]. Dominant features were selected using GBDT and incremental feature selection method. The obtained 616 features were used to train and test SVM model and it achieved 75.7% accuracy.

By considering the dataset of different species protein-protein interaction can be predicted has been discussed in reference [16]. The comparative analysis is done on Human, S. cerevisiae, E. coli, C. elegans, H. pylori and M. musculus datasets. Based on the physiochemical properties of amino acid, the features were extracted using dipeptide composition methods. 5-layer fully connected DNN is built with the activation function ReLu and the PPI is predicted for 256 features with 60567 instances. For the benchmark dataset 99.57% accuracy is achieved.

Prediction of human-virus protein-protein interaction is carried out using various machine learning techniques has been discussed in reference [17]. The positive and negative samples taken from HPIDB are treated with CD-HIT to remove the redundancy. The retrieved 291,726 sequences were used for training. The unsupervised doc2vec embedding learning model is used to extract additional features of protein sequence has been discussed in references [17] and [32]. The performance of RF, SVM and Adaboost models were analysed with 20% and 80% testing and training sets respectively. Random forest model achieved 93.23% accuracy for doc2vec vector features.

The random Forest algorithm performs well with mRMR feature selection method to predict the PPI has been discussed in reference [18]. 25,856 instances were selected after removing redundant and homologous sequences. By considering all the physiochemical properties of proteins, their features are extracted and prominent features was selected using mRMR method [19]–[33]. The obtained 51 features were used for training and testing the RF model. Obtained accuracy is 67.29%.

The performance of different algorithms and the dataset details is presented in Table 1. The experimental procedure for detecting PPIs takes a long time and are costly. SVM and

KNN models can be used to predict protein interactions. The binary coding method can be used to construct the features of 3,271 binary interactions. 2,338 proteins provide a prediction accuracy of 98,11% for KNN classifier and 86.99% for SVM model. FIGURE 1 exhibits the accuracy obtained for different ML algorithms when different datasets are used.

TABLE 1. Performance comparison of different algorithms to predict API.

Paper	FS method and Algorithm	Dataset	Accuracy
Emami N etal[11]	RF and DNN	3404x640	91.38%
W. Lee etal[12]	RF and SVM	38327x10	97.76%
Zhang etal.[13]	RIFS and Ensemble Classifier	2900x304	73.2%
Li BQ etal,[14]	mRMR and RFA	2900x290	77.41%
Qing Yang etal15]	GBDT and RFA	2900x616	75.7%
Lei Yang etal[16]	DC and DNN	60567x256	99.57%
Xiaodi Yang etal[17]	Doc2vecLM and RFA	291,726seq	93.23%
Li BQ etal [18]	mRMR and RFA	25,856x51	67.29%
You Z etal[20][21]	SVM and KNN	2338seq	98%

III. METHODLOGY

Aptamers are molecules that bend towards specific structures and attach to certain receptors like proteins. They usually restrict protein–protein interactions in related sites, which can result in medicinal properties like antagonism. Identifying the aptamer-protein interaction pairs with the aid of machine learning reduces the clinical process costs. FIGURE 2 shows the block diagram of the proposed model.

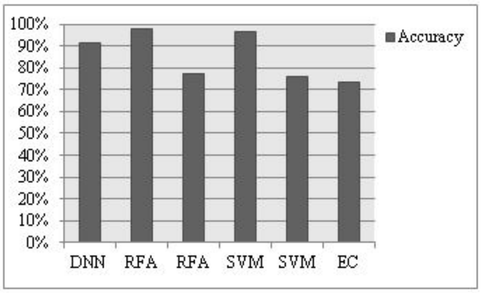


FIGURE 1. Performance of ML algorithms in API prediction.

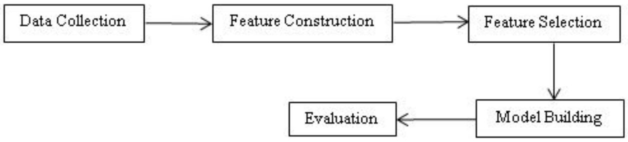


FIGURE 2. Block diagram of the proposed model.

A. DATA COLLECTION

ApatmerBase [22] is a community source of information regarding oligonucleotides, including different experimental settings and reference to main scientific literature has been discussed in reference [23]. 1638 records were found from the AptamerBase out of which 257 were target proteins and 1638 were aptamers. Sequences of those entries were collected from NCBI and RCSB website for the best matching Id. Molecular docking is performed by arbitrarily making paring of the proteins and aptamers. Finally 2900 instances were considered with 2175 negative and 725 positive instances. Therefore 2900 instances were used for feature construction.

B. FEATURE CONSTRUCTION

In this section, PseudoAminoAcidComposition (PseAAC), AminoAcidCompoistio (AAC) and NucleotideComposition (NC) are discussed in detail. The target protein sequences were encoded using amino acid composition and pseudo amino acid composition method, whereas the aptamer sequences were encoded using nucleotide composition.

1) NUCLEOTIDE COMPOSITION (NC)

The four base pairs of DNA and RNA are ATGC and AUGC respectively. Every aptamer was represented numerically as a 20-dimensional variable by using the Nucleotide Composition Method.

2) AMINO ACID COMPOSITION (AAC)

It computes the number of occurrences of amino acid in a given residue with the Eq-1 where T is the type of amino acid and M is the total number of amino acid in the residue.

$$\mathrm{Freq(T)} = \mathrm{(M(T))/M} \tag{1}$$

3) PSEUDO AMINO ACID COMPOSITION (PSEAAC)

It is used to extract features from the sequences constructively. If PR is the protein with the chain M then its residues can be represented with the Eq-2.

$$\mathrm{PR = Rs1, Rs2, Rs3,..up\ to\ RsM} \tag{2}$$

Sequence order effect can be computed by a collection of distinct correlation factors [12] as shown in Eq-3. The correlation factors are $\theta 1, \theta 2, \theta 3 \ldots, \theta\lambda$

$$\begin{aligned}
\theta 1 &= (\frac{1}{l1-1})\sum_{i=1}^{l1-1}\theta(Rsi, Rsi+1)\\
\theta 2 &= (\frac{1}{l1-2})\sum_{i=1}^{l1-2}\theta(Rsi, Rsi+2)\\
\theta\lambda &= (\frac{1}{l1-\lambda})\sum_{i=1}^{l1-\lambda}(Rsi, Rsi+\lambda)
\end{aligned} \tag{3}$$

The correlation function can be computed by using Eq-4. P1(Rsi), P2(Rsj), Ms(Rsi) etc., are the physiochemical properties of amino acids and Eq-5 is used to change each of these values from their original values P1(i), P2(i), Ms(i).

$$\begin{aligned}
\theta(Rsi, Rsj) &= (\frac{1}{3})[[\mathrm{P1(Rsj)\text{-}P1(Rsi)}]^2 + [P2(Rsj)\\
&\quad - P2(Rsi)]^2 + [Ms(Rsj) - Ms(Rsi)]^2]
\end{aligned} \tag{4}$$

$$P1(i) = \frac{P1(i) - \sum i = 1^{20} P1(i)/20}{\sqrt{\sum i = 1^{20}[P1(i) - \sum i = 1^{20} P1(i)/20}}]$$
$$P2(i) = \frac{P2(i) - \sum i = 1^{20} P2(i)/20}{\sqrt{\sum i = 1^{20}[P2(i) - \sum i = 1^{20} P2(i)/20}}]$$
$$Ms(i) = \frac{Ms(i) - \sum i = 1^{20} Ms(i)/20}{\sqrt{\sum i = 1^{20}[Ms(i) - \sum i = 1^{20} Ms(i)/20}}] \tag{5}$$

Finally, a vector vec1, vec2.. May be used to depict the protein PR's PseAAC with $20 + \lambda$ dimensions as shown in Eq-6

$$[\text{vec1, vec2, vec3} \ldots\ldots \text{vec20} + \lambda]^{T}\text{rWhere, } T_{r}\text{transpose} \tag{6}$$

We have considered 18-Physiochemical properties and it is collected from references [24] and [25]. Properties such as bulkiness, buriability, molecular weight, melting point, hydrophobicity, unfolding entropy, enthalpy and gibs free energy, polarity, sidechain mass, volume of residue, etc., are used in our work [24], [25].

C. FEATURE SELECTION

Selection of features is a fundamental topic in learning algorithms that have a significant influence on the model's efficacy [30]. The attributes we use for training the models do have significant impact on the results we get. Prediction accuracy can be harmed by unrelated attributes. The feature selection process reduces the error rate, improves performance and training time is cut in half has been discussed in reference [31]. After performing PseAAC method 290 features were extracted for 2900 instances. Compared to KNN, XGB and SVM, Random forest model yielded 80% accuracy for all 290 features and it is shown in FIGURE 3.

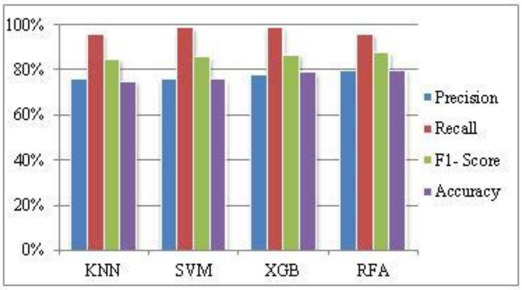


FIGURE 3. Performance of ML algorithms for all 290 features.

Methods that give a rank to input characteristics depending on how valuable they are at detecting a class label are known as feature importance. In order to analyse the result RFFI and XGBFI are used. Both the decision tree feature importance model give the property feature_importances_ and it is used to get the relative relevance ratings to every attribute in the input. FIGURE 3 shows the top 20 and FIGURE 4 shows the top 23 features based on its score using RFFI and XGBFI method respectively.

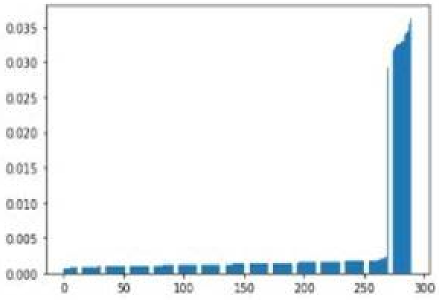


FIGURE 4. Selection of features using RFFI method.

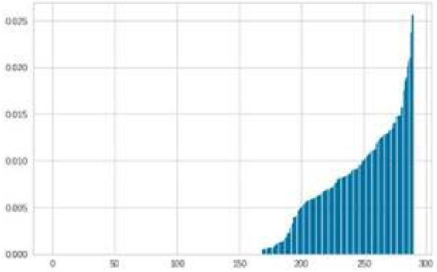


FIGURE 5. Selection of features using XGBFI method.

One of the common dimensionality reduction methods is principal component analysis. It aids there in the extraction of a vector of features from a huge number of existing components. Principal Components (PC) are the key parameters that have been retrieved and are combined in a linear fashion. The very first PC is retrieved in such a manner that it describes the highest variance there in the data source. The second PC, which has nothing to do with the very first, attempts to describe the exceptional variance in data sources. The third PC, which has nothing to do with the second, attempts to describe the exceptional variance in data sources and so on.

While performing PCA, we are allowed to randomly mention the number of attributes (n) which we want to retain. Out of 290 features, we have done this experiment to retain principal 20, 25, 50, 100, 150 and 200 features to analyse the performance of various machine learning models. For the mentioned attributes variance will be calculated and it keeps adding the attributes until a cutoff point. Later, it sorts the characteristics according to the amount of variance they represent and then plot the cumulative proportion of variance for the first n components.FIGURE 6, FIGURE 7 and FIGURE 8 shows the principal 20, 25 and 50 components, respectively.

FIGURE 9, FIGURE 10 and FIGURE 11 shows the principal 100, 150 and 200 components respectively.

D. MODEL BUILDING

Performance analysis is carried out for the various machine learning models choosing the different set of features which is obtained after performing RFFI, XGBFI and PCA. The dominant 20 features which is retrieved after performing RFFI is considered along with 2900 instances and 75% of the dataset was used as training set and the rest 25% as testing set.

SVM, KNN, XGB and RF models were built and trained with the training dataset. The type of kernel used in SVM is linear and when n_neighbors of KNN is set to 8 it yielded 75% accuracy. Result of XGB and RF was good when n_estimators was set to 60 and 8 respectively. The Fig.12 shows the performance of ML models for 20 features and 2900 instances. SVM, XGB and KNN models yielded 75% accuracy.

The dominant 23 features which is retrieved after performing XGBFI was considered along with 2900 instances and 75% of the dataset was used as training set and the rest 25% as testing set. The FIGURE 13 shows the performance of ML models for 23 features and 2900 instances. RF model yielded 72% accuracy, whereas the other three models yielded 75% accuracy.

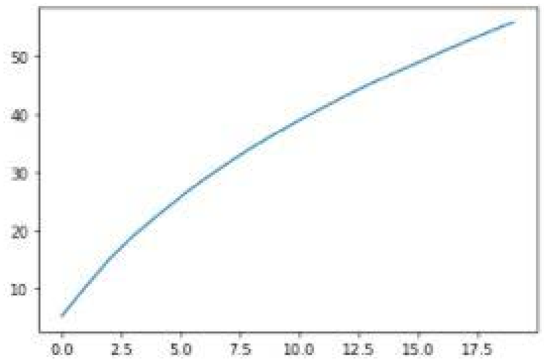


FIGURE 6. Principal 50 components.

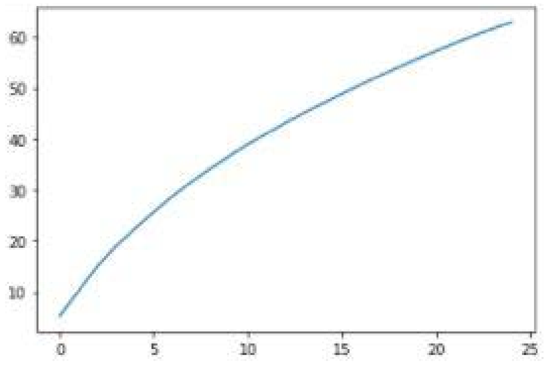


FIGURE 7. Principal 200 components.

Computational time will be reduced when the dominant features are selected and used for the prediction. The performance of the model is also based on the number of attributes which support the instances. After performing a PCA different set of principal features were selected and the same has been used for prediction. The obtained 20, 25, 50,100,150 and 200 features were considered along with 2900 instances. KNN, SVM and XGB will not much show any changes in its result when principal 100,150 and 200 features are employed, but the accuracy of the RF model will decrease. Hence, RF model gives best result for principal 50 features with 2900 instances when n_estimators is set to 7.

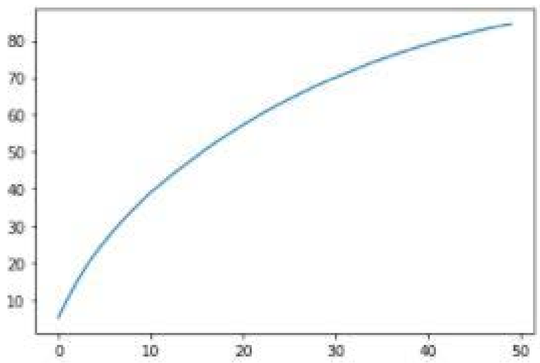


FIGURE 8. Classifiers Vs Accuracy for the prominent features selected through RFFI method.

KNN, SVM and XGB will not much show any changes in its result when principal 100,150 and 200 features are employed, but the accuracy of the RF model will decrease. Hence, RF model gives best result for principal 50 features with 2900 instances when n_estimators is set to 7. FIGURE 17, FIGURE 18, and FIGURE 19 show the result of the classifiers when 100, 150 and 200 principal features are selected.

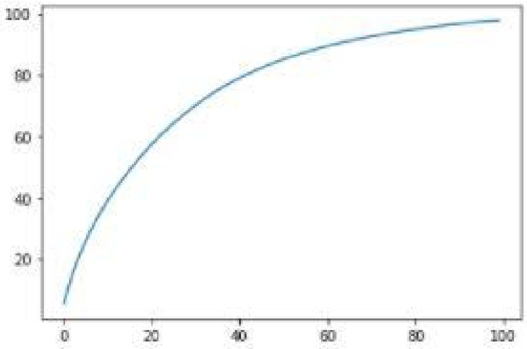


FIGURE 9. Classifiers Vs Accuracy for the prominent features selected through XGBFI method.

E. EVALUATION

The confusion matrix helps determine model's correctness. Precision, Recall, F1-Score and Accuracy can be calculated by using Eq-7, Eq-8, Eq-9 and Eq-10 respectively.

$$Precision = \frac{TrPos}{TrPos + FalPos} \tag{7}$$

$$Recall = \frac{TrPos}{TrPos + FalNeg} \tag{8}$$

$$F1\_Score = 2 * \frac{Precison * Recall}{Precision + Recall} \tag{9}$$

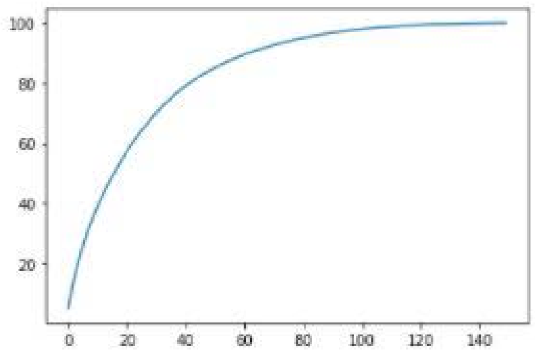


FIGURE 10. Classifiers Vs Accuracy for 50 principal components.

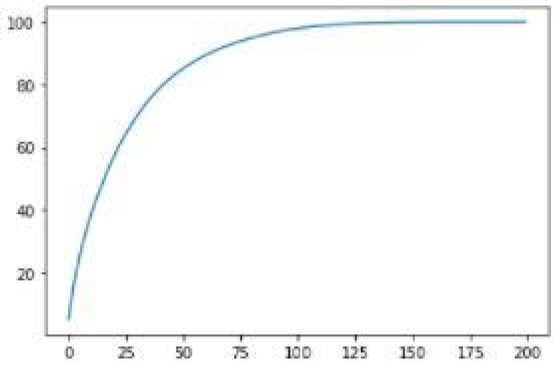


FIGURE 11. Classifiers Vs Accuracy for 200 principal components.

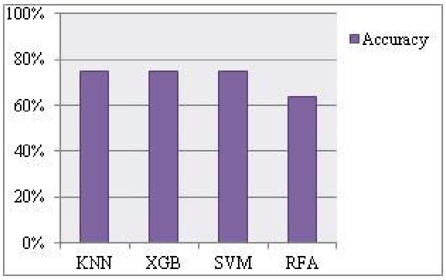


FIGURE 12. Principal Components Vs Accuracy of RFA model.

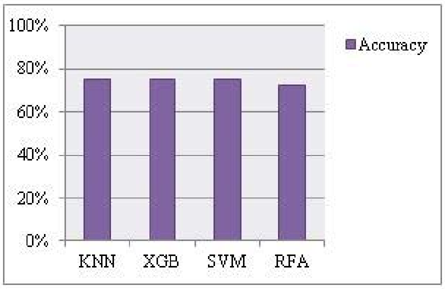


FIGURE 13. Accuracy Vs n_estimators of RFA.

$$Accuracy = \frac{TrN + TrP}{TrN + TrP + FalP + FalN} \tag{10}$$

KNN, SVM, and XGB models yielded 75% accuracy for top 20 and 23 features after performing RFFI and XGBFI respectively. Features were retrieved based on their score. The performance of ML algorithms was analysed when those features were considered with the 2900 instances. The performance of classifiers is as shown in Table 2 and Table 3.

TABLE 2. Performance of classifiers for selected features using RFFI.

Classifiers	Precision	Recall	F1 Score	Accuracy
KNN	75%	100%	86%	75%
SVM	75%	100%	86%	75%
XGB	75%	100%	85%	75%
RFA	72%	86%	75%	64%

TABLE 3. Performance of classifiers for selected features using XGBFI.

Classifiers	Precision	Recall	F1 Score	Accuracy
KNN	75%	100%	86%	75%
SVM	75%	100%	86%	75%
XGB	75%	100%	85%	75%
RFA	75%	95%	84%	72%

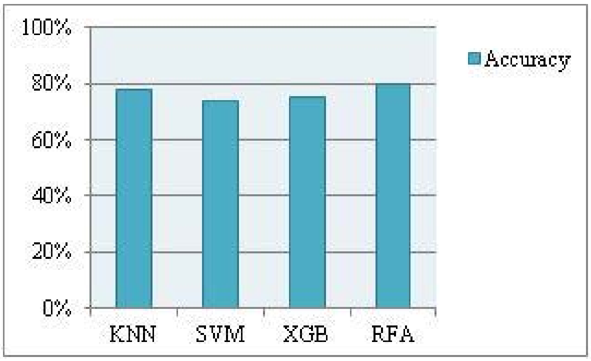


FIGURE 14. Classifiers Vs Accuracy for 20 principal components.

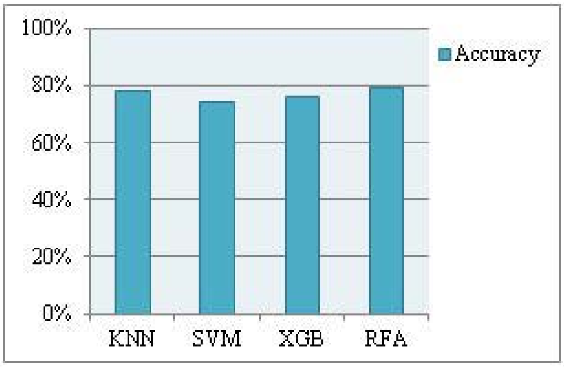


FIGURE 15. Classifiers Vs Accuracy for 25 principal components.

With in testing set, KNN finds the distance among a demand and then all the instances in the dataset, selects the given lot of instances (K) nearest towards the enquiry, and afterwards decides for its most common label. In the case of classification and regression, we found that the best way to

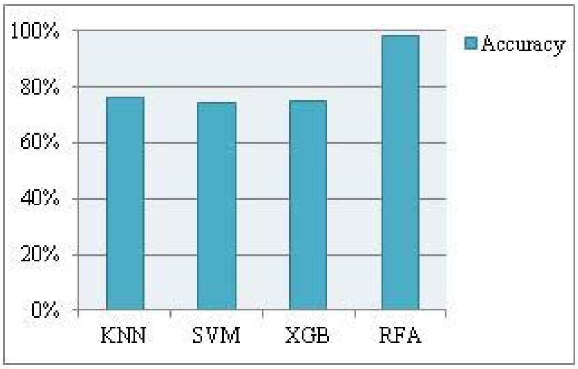


FIGURE 16. Classifiers Vs Accuracy for 50 principal components.

TABLE 4. Performance of KNN for principal components.

PCs	Precision	Recall	F1 Score	Accuracy
20	**79%**	**95%**	**86%**	**78%**
25	79%	94%	86%	78%
50	80%	90%	85%	76%
100	78%	94%	85%	76%
150	78%	94%	85%	75%
200	76%	97%	85%	75%

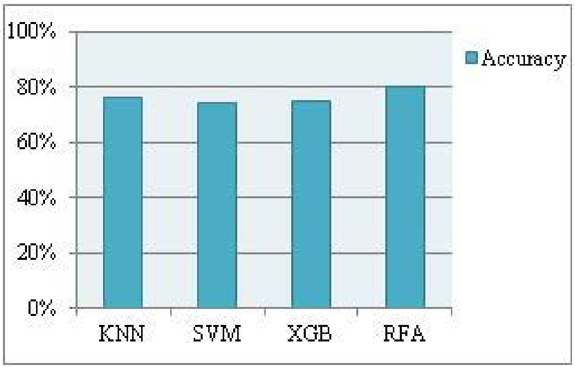


FIGURE 17. Classifiers Vs Accuracy for 100 principal components.

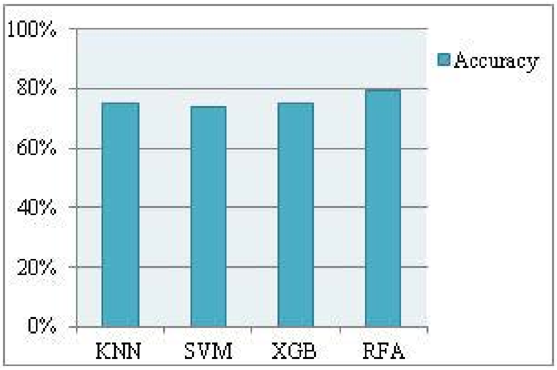


FIGURE 18. Classifiers Vs Accuracy for 150 principal components.

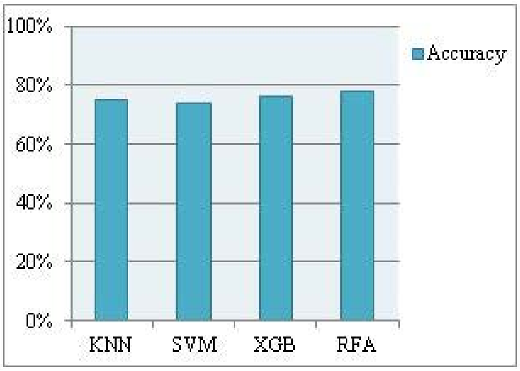


FIGURE 19. Classifiers Vs Accuracy for 200 principal components.

choose the correct K for our data is to try a few different Ks and see which one performs best. Performance of KNN for the chosen principal components depicted in Table 4. In our work, KNN model has given a maximum of 78% accuracy for 20 principal components. The Support Vector Machine, is a linear classifier that may be used to solve regression and classification tasks. It can handle both linear and nonlinear problems and is useful for a wide range of applications. SVM is a basic concept: The method split the entire dataset into categories by drawing a line or hyperplane. Performance of SVM for the chosen principal components depicted in Table 5. In our work, SVM model has given a maximum of 74% accuracy for 20 principal components.

TABLE 5. Performance of SVM for principal components.

PCs	Precision	Recall	F1 Score	Accuracy
20	**74%**	**100%**	**85%**	**74%**
25	74%	100%	85%	74%
50	74%	100%	85%	74%
100	74%	100%	85%	74%
150	74%	99%	85%	74%
200	74%	99%	85%	74%

TABLE 6. Performance of SVM for different kernel functions at each principal components.

PCs	RBF	Sigmoid	Plynomial	Linear
20	74%	64%	74%	74%
25	74%	73%	82%	74%
50	75%	73%	76%	74%

SVM techniques are based on a collection of mathematical functions known as the kernel. The kernel's job is to take data and turn it into the needed format[28], [29]. Table 6 shows the performance of SVM for different kernel functions for the principal components.

XGBoost is a method of ensemble learning. It may not always be enough to depend on the findings of a single machine learning model. Ensemble learning is a method for combining the predictive capacity of several learners in a systematic way. The end result is a single model that combines the outputs of many models. Performance of XGBoost for the chosen principal components depicted in Table 7. In our work, XGB model has given a maximum of 76% accuracy for 25 principal components.

TABLE 7. Performance of XGB for principal components.

PCs	Precision	Recall	F1 Score	Accuracy
20	74%	100%	85%	75%
25	**75%**	**99%**	**86%**	**76%**
50	75%	100%	86%	75%
100	75%	100%	86%	75%
150	75%	100%	86%	75%
200	76%	99%	86%	76%

TABLE 8. Performance of RFA for principal components.

PCs	Precision	Recall	F1 Score	Accuracy
20	81%	94%	87%	80%
25	80%	92%	86%	79%
50	**97%**	**100%**	**98%**	**98%**
100	80%	98%	88%	80%
150	79%	98%	87%	79%
200	78%	98%	87%	78%

Random forests is a method for supervised learning. It has the ability to be utilised for both classification and regression. It's also the most adaptable and user-friendly algorithm. The trees make up a forest. A forest is believed to be more strong the more trees it has. RF constructs decision trees from arbitrarily chosen samples, receives predictions from each tree, and votes on the right answer. Performance of RFA for the chosen principal components depicted in Table 8. In our work, RF model has given a maximum of 98% accuracy for 50 principal components.

Kernel PCA projects a sample into a greater feature map, where it may be linearly separated, using a kernel function. The Table 9 and Table 10 shows the performance of KNN and GB respectively for 20,25,50,100,150 and 200 principal components. Both the models yields good result with 50 principal components as shown depicted.

TABLE 9. Performance of KNN after applying KernelPCA.

PCs	Precision	Recall	F1 Score	Accuracy
20	79%	94%	86%	77%
25	**79%**	**95%**	**86%**	**78%**
50	80%	90%	85%	76%
100	78%	92%	85%	75%
150	79%	94%	86%	77%
200	76%	97%	85%	75%

TABLE 10. Performance of XGB after applying KernelPCA.

PCs	Precision	Recall	F1 Score	Accuracy
20	74%	100%	85%	74%
25	75%	100%	86%	75%
50	**76%**	**100%**	**86%**	**76%**
100	76%	100%	86%	76%
150	76%	99%	86%	76%
200	76%	99%	86%	76%

KNN and GB model yields good performance with principal 25 and 50 components respectively. Table 11 and Table 12 shows the result of SVM and RFA respectively for the principal components. Bothe the models yields good result with 50 principal components.

TABLE 11. Performance of SVM after applying KernelPCA.

PCs	Precision	Recall	F1 Score	Accuracy
20	75%	100%	85%	75%
25	77%	98%	86%	77%
50	**79%**	**96%**	**87%**	**78%**
100	74%	100%	85%	74%
150	74%	99%	85%	74%
200	74%	99%	85%	74%

TABLE 12. Performance of RFA after applying KernelPCA.

PCs	Precision	Recall	F1 Score	Accuracy
20	81%	93%	87%	79%
25	82%	93%	87%	79%
50	**97%**	**100%**	**98%**	**97%**
100	81%	97%	88%	81%
150	81%	97%	88%	81%
200	78%	98%	87%	78%

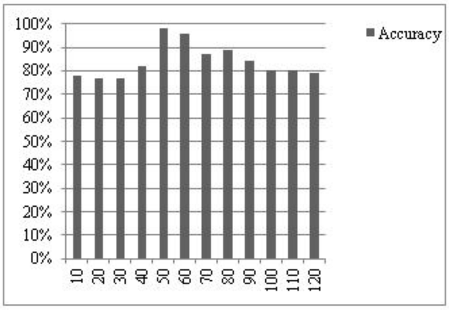


FIGURE 20. Principal Components Vs Accuracy of RFA model.

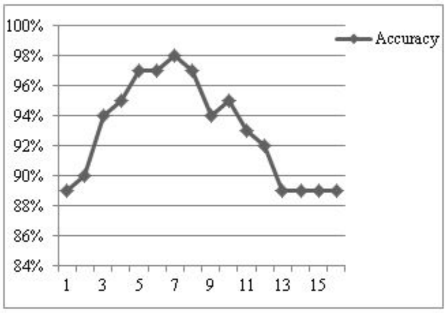


FIGURE 21. Accuracy Vs n_estimators of RFA.

IV. RESULTS AND DISCUSSION

Sequences of the aptamers and proteins were collected from the NCBI website. After the docking study, we got 2900 instances with 2175 negative and 725 positive instances. These were used for future construction and it is carried out by PseAAC method. We have used 18 physiochemical properties to compute correlation function as a result

290 features were constructed. When the performance of the classifiers is analysed for 290 features, we observed that the RF model gave 80% accuracy compared to KNN, SVM and XGB. Feature selection process plays an important role and it makes the model to perform well. XGBFI and RFFI are the feature importance classifiers which calculate the ranks or scores for each feature. Based on the XGFI score, dominant 23 dimensions were selected out of 290 and used along with 2900 instances to analyse the result. The same experiment is conducted with RFFI and we got 20 features. We achieved maximum of 75% accuracy for all the three models i.e. KNN, SVM and XGB in both the methods; results are shown in Table 2.

One of the common dimensionality reduction methods is principal component analysis. It aids there in the extraction of a vector of features from a huge number of existing components. While performing PCA, we are allowed to randomly mention the number of attributes (n) which we want to retrieve. When this experiment is carried out we observed that Random forest gives 98% accuracy for 50 principal components. The FIGURE 20 shows the result of RF model for various principal components and FIGURE 21 shows the accuracy with respect to the n_estimators.

V. CONCLUSION

Interactions between aptamers and proteins are significant in physiological activities and molecular identification. Our method took into account not only the genetic material from oligonucleotides, but also the conventional and the pseudo amino acid composition of target protein. With the help of dimensionality reduction method, principal components were selected and performance evaluation was carried out. We achieved 98% accuracy for the Random Forest model for 50 principal components. These attributes may provide guidance for building unique and effective aptamers that bind to specific targets, allowing for a better understanding of the mechanics of interaction among aptamers and their target.

REFERENCES

[1] K.-M. Song, S. Lee, and C. Ban, "Aptamers and their biological applications," *Sensors*, vol. 12, no. 1, pp. 612–631, Jan. 2012.

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

[3] M. Vidic, T. Smuc, N. Janez, M. Blank, T. Accetto, J. Mavri, I. C. Nascimento, A. A. Nery, H. Ulrich, and T. T. Lah, "In silico selection approach to develop DNA aptamers for a stem-like cell subpopulation of non-small lung cancer adenocarcinoma cell line A549," *Radiol. Oncol.*, vol. 52, no. 2, pp. 152–159, Mar. 2018.

[4] J. He, J. Wang, N. Zhang, L. Shen, L. Wang, X. Xiao, Y. Wang, T. Bing, X. Liu, S. Li, and D. Shangguan, "*In vitro* selection of DNA aptamers recognizing drug-resistant ovarian cancer by cell-SELEX," *Talanta*, vol. 194, pp. 437–445, Mar. 2019.

[5] W.-P. Hu, J. V. Kumar, C.-J. Huang, and W.-Y. Chen, "Computational selection of RNA aptamer against Angiopoietin-2 and experimental evaluation," *BioMed Res. Int.*, vol. 2015, pp. 1–8, Mar. 2015.

[6] D. S. Shcherbinin, O. V. Gnedenko, S. A. Khmeleva, S. A. Usanov, A. A. Gilep, A. V. Yantsevich, T. V. Shkel, I. V. Yushkevich, S. P. Radko, A. S. Ivanov, A. V. Veselovsky, and A. I. Archakov, "Computeraided design of aptamers for cytochrome p450," *J. Struct. Biol.*, vol. 191, no. 2, pp. 112–119, 2015.

[7] R. Ahirwar, S. Nahar, S. Aggarwal, S. Ramachandran, S. Maiti, and P. Nahar, "In silico selection of an aptamer to estrogen receptor alpha using computational docking employing estrogen response elements as aptamer-alike molecules," *Sci. Rep.*, vol. 6, no. 1, Aug. 2016, Art. no. 21285.

[8] X. Zhou, B. Park, D. Choi, and K. Han, "A generalized approach to predicting protein-protein interactions between virus and host," *BMC Genomics*, vol. 19, no. S6, p. 568, Aug. 2018.

[9] B. Kim, S. Alguwaizani, X. Zhou, D.-S. Huang, B. Park, and K. Han, "An improved method for predicting interactions between virus and human proteins," *J. Bioinf. Comput. Biol.*, vol. 15, no. 1, Feb. 2017, Art. no. 1650024.

[10] W. Lee and K. Han, "Constructive prediction of potential RNA aptamers for a protein target," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 5, pp. 1476–1482, Sep. 2020.

[11] N. Emami and R. Ferdousi, "AptaNet as a deep learning approach for aptamer–protein interaction prediction," *Sci. Rep.*, vol. 11, no. 1, p. 6074, Dec. 2021.

[12] W. Lee and K. Han, "Constructive prediction of potential RNA aptamers for a protein target," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 5, pp. 1476–1482, Sep. 2020.

[13] L. Zhang, C. Zhang, R. Gao, R. Yang, and Q. Song, "Prediction of aptamer-protein interacting pairs using an ensemble classifier in combination with various protein sequence attributes," *BMC Bioinf.*, vol. 17, no. 1, p. 225, Dec. 2016.

[14] B. Q. Li, Y. C. Zhang, G. H. Huang, W. R. Cui, N. Zhang, and Y. D. Cai, "Prediction of aptamer-target interacting pairs with pseudo-amino acid composition," *PLOS ONE*, vol. 9, no. 1, 2014, Art. no. e86729.

[15] Q. Yang, C. Jia, and T. Li, "Prediction of aptamer–protein interacting pairs based on sparse autoencoder feature extraction and an ensemble classifier," *Math. Biosciences*, vol. 311, pp. 103–108, May 2019.

[16] L. Yang, Y. Han, H. Zhang, W. Li, and Y. Dai, "Prediction of protein-protein interactions with local weight-sharing mechanism in deep learning," *BioMed Res. Int.*, vol. 2020, pp. 1–11, Jun. 2020.

[17] X. Yang, S. Yang, Q. Li, S. Wuchty, and Z. Zhang, "Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 153–161, Jan. 2020.

[18] B.-Q. Li, K.-Y. Feng, L. Chen, T. Huang, and Y.-D. Cai, "Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS," *PLoS ONE*, vol. 7, no. 8, Aug. 2012, Art. no. e43927.

[19] X.-W. Chen and M. Liu, "Prediction of protein-protein interactions using random decision forest framework," *Bioinformatics*, vol. 21, no. 24, pp. 4394–4400, Dec. 2005.

[20] Z. You, Z. Ming, B. Niu, S. Deng, and Z. Zhu, "A SVM-based system for predicting protein-protein interactions using a novel representation of protein sequences," in *Proc. Int. Conf. Intell. Comput.*, vol. 7995, D. S. Huang, V. Bevilacqua, J. C. Figueroa, and P. Premaratne, Eds. 2013, pp. 629–637.

[21] M. R. Guarracino and A. Nebbia, "Predicting protein-protein interactions with K-nearest neighbors classification algorithm," in *Proc. Int. Meeting Comput. Intell. Methods Bioinf. Biostatistics* in Lecture Notes in Computer Science, vol. 6160, R. L. E. Tagliaferri, Eds. 2010, pp. 139–150.

[22] *Aptamer and Protein IDS*. Accessed: Jun. 2012. [Online]. Available: http://aptamer.freebase.com/

[23] J. Cruz-Toledo, M. McKeague, X. Zhang, A. Giamberardino, and E. McConnell, "Aptamer Base: A collaborative knowledge base to describe aptamers and SELEX experiments," *Database*, vol. 2012, Jan. 2012, Art. no. bas006.

[24] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: Amino acid index database, progress report 2008," *Nucleic Acids Res.*, vol. 36, pp. D202–D205, Dec. 2007.

[25] M. M. Gromiha, "A statistical model for predicting protein folding rates from amino acid sequence with structural class information," *J. Chem. Inf. Model.*, vol. 45, no. 2, pp. 494–501, Mar. 2005.

[26] H. L. Gururaj, F. Flammini, and H. A. C. Kumari, "Classification of drugs based on mechanism of action using machine learning techniques," *Discover Artif. Intell.*, vol. 1, pp. 1–14, Dec. 2021, doi: 10.1007/s44163-021-00012-2.

[27] C. D. Divya, H. L. Gururaj, R. Rohan, V. Bhagyalakshmi, H. A. Rashmi, A. Domnick, and F. Flammini, "An efficient machine learning approach to nephrology through iris recognition," *Discover Artif. Intell.*, vol. 1, pp. 1–15, Dec. 2021, doi: 10.1007/s44163-021-00010-4.

[28] V. Dave, S. Singh, and V. Vakharia, "Diagnosis of bearing faults using multi fusion signal processing techniques and mutual information," *Indian J. Eng. Mater. Sci.*, vol. 27, no. 4, pp. 878–888, 2020.

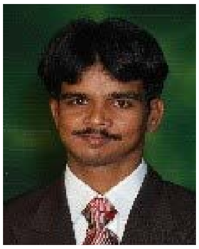
[29] Y. Liu, J. Lian, M. R. Bartolacci, and Q.-A. Zeng, ''Density-based penalty parameter optimization on C-SVM,'' *Sci. World J.*, vol. 2014, pp. 1–9, Jan. 2014, doi: 10.1155/2014/851814.

[30] V. Vakharia, V. K. Gupta, and P. K. Kankar, ''A comparison of feature ranking techniques for fault diagnosis of ball bearing,'' *Soft Comput.*, vol. 20, no. 4, pp. 1601–1619, Apr. 2016, doi: 10.1007/s00500-015-1608-6.

[31] R.-C. Chen, C. Dewi, S.-W. Huang, and R. E. Caraka, ''Selecting critical features for data classification based on machine learning methods,'' *J. Big Data*, vol. 7, no. 1, p. 52, Dec. 2020, doi: 10.1186/s40537-020-00327-4.

[32] H. L. Gururaj, F. Flammini, H. A. C. Kumari, G. R. Puneeth, and B. R. S. Kumar, ''Classification of drugs based on mechanism of action using machine learning techniques,'' *Discover Artif. Intell.*, vol. 1, no. 1, pp. 1–14, Dec. 2021, doi: 10.1007/s44163-021-00012-2.

[33] C. D. Divya, H. L. Gururaj, and R. Rohan, ''An efficient machine learning approach to nephrology through iris recognition,'' *Discover Artif. Intell.*, vol. 1, pp. 1-15, Dec. 2021, doi: 10.1007/s44163-021-00010-4.



N. MANJU received the B.E. degree in computer science and engineering, the M.Tech. degree in computer network engineering, and the Ph.D. degree from Visvesvaraya Technological University, Belagavi, Karnataka, India, in 2005, 2009, and 2021, respectively. He is currently working as an Assistant Professor with the Department of Information Science and Engineering, Sri Jayachamarajendra College of Engineering, Mysore, Karnataka. His research interests include machine learning, computational intelligence, and computer networks. He is a Life Member of ISTE. He was also a reviewer of international journals and conferences.

C. M. SAMIHA received the M.Tech. degree from JSS Science and Technological University, Mysore. She is currently working as an Assistant Professor with the Department of Computer Science. She has eight years of teaching experience at UG level. Het research interests include bioinformatics and machine learning.

S. P. PAVAN KUMAR is currently pursuing the Ph.D. degree in computer science and engineering from Visvesvaraya Technological University, Belagavi. He has eight years of teaching and research experience at both UG and PG level. He is currently working as an Assistant Professor with the Department of Computer Science and Engineering. He has published more than ten research articles in peer-reviewed journals. He has presented more than five papers at various international conferences. He worked as a reviewer for various journals and conferences. He trained more than 500 professionals on python programming and machine learning. His research interests include bioinformatics, computational biology, the IoT, artificial intelligence, and machine learning. He is a Professional Member of IEEE. He is also an EXECOM Member of IEEE Mysuru Subsection. He has honored as a resource person at seminars and workshops.

H. L. GURURAJ (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from Visvesvaraya Technological University, Belagavi, in 2019. He has ten years of teaching and research experience at both UG and PG levels. He is currently working as an Associate Professor with the Department of Computer Science and Engineering. He is the Founder of Wireless Internetworking Group (WiNG). He is a Professional Member of ACM. He is a Senior Member of ACM, a Lifetime Member of the Cryptology Society of India, and a Senior Member of the IEEE Computer Society. He was appointed as an ACM Distinguish Speaker (2018–2021) by the ACM U.S. Council and he is one among 15 speakers across India. He received the Young Scientist Award from ITS-SERB, Department of Science and Technology, Government of India, in December 2016. He also received the Best Project Guide Award consecutively for four years at the Malnad College of Engineering, India, in 2015, 2016, and 2017. He worked as a Special Editor of EAI publisher and a Guest Editor of Multimedia System (Springer). He is an editorial board member of various international journals.

FRANCESCO FLAMMINI (Senior Member, IEEE) received the master's *(cum laude)* and Ph.D. degrees in computer engineering from the University of Naples Federico II, Italy, in 2003 and 2006, respectively. He has worked for 15 years in private and public companies, including Ansaldo STS (now Hitachi Rail) and Italian State Mint and Polygraphic Institute (IPZS), on large international projects addressing intelligent systems and cybersecurity, as a Technical Leader and the Unit Head. He has been a Senior Lecturer and the Chair of the Cyber-Physical Systems Group, Linnaeus University, Sweden; and a Full Professor of computer science with a focus on cyber-physical systems at Mälardalen University, Sweden. He has (co)authored more than 150 scientific publications. His research interest includes trustworthy autonomy. He is an ACM Distinguished Speaker. He has served as the chair, an invited speaker, a steering/program committee member, and an editor for several international conferences and journals.