Chapter 10

Development of web application for the identification of anti-Alzheimer's ligands

10. Development of web application for identification of anti-Alzheimer's ligands

10.1. Introduction

AD is a neurodegenerative disease with the involvement of various enzymes and receptors that could be therapeutic targets. The various hypotheses regarding the pathology and progression of AD include amyloid-β cascade, tau hypothesis, oxidative stress, neuroinflammation, glucose hypermetabolism, cholinergic disruption and loss, gut microbiome, bacteria-derived metabolites, some immune and endocrinal related pathways [4]. Some of the important therapeutic targets under clinical investigation are BACE1, GSK- 3β, MAO- B, matrix metalloproteases, NMDA receptors, tau kinase [246]. The present study deals with the development of a web application to identify inhibitors of AChE, BChE, BACE1, GSK-3β, MAO-B and N2B subunit of NMDA receptors. The datasets for the targets were obtained, processed and molecular descriptors were calculated. The descriptors were used for training the ML models using various binary classification algorithms and were validated. The models that performed well were selected and deployed in a web application for the users.

10.2. Materials and methods

10.2.1. Dataset preparation

The datasets of the inhibitors of the selected targets were obtained from the Bindingdb database (http://bindingdb.org/bind/index.jsp) [270]. The datasets were preprocessed to remove compounds with missing IC_{50} values, nonstandard structures and duplications. The organophosphorus compounds were removed in the case of AChE and BChE.

10.2.2. Molecular descriptor calculation

Molecular descriptors were calculated by employing the RDkit module in python for each dataset [283]. A total of 208 descriptors were calculated from the SMILES string of the molecules.

10.2.3. Classification of the datasets

The IC_{50} values of the inhibitors of a dataset were converted into two classes, i.e., active (IC_{50} <= 5000 nM) and inactive (IC_{50} > 5000 nM) for the development of binary classification models. **Figure 10.1** represent the division of various datasets in binary classes.

10.2.4. Feature selection

The 208 descriptors calculated for each dataset were further reduced in order to improve the accuracy of the ML model as well as to decrease the computational time. Initially, a zero-variance filter was applied to drop a feature with zero variance. It was followed by the application of the correlation filter with the criteria that only one of the features would be used in a group of features that show a correlation of more than 0.9. Finally, a sequential feature selector was applied that used an RF classifier to obtain the feature importance of individual descriptors in a dataset. This resulted in the preparation of the final descriptors set with 25 features each for all the targets. **Figure 10.2** represent the various stages of the feature selection process.

10.2.5. Division of dataset

Each dataset was divided into three groups, i.e., training, validation and test set in a ratio of 70:15:15. The training set was deployed for model training, while the validation and test sets were used for determining the accuracy of the model.

10.2.6. Training of machine learning models

Six binary classification algorithms viz. LR, KNN, SVC, Bernoulli Naïve Bayes, decision tree and RF classifier were employed to train models on the training dataset. The LR, KNN and SVC, were trained on a standardised dataset while others were trained on the actual dataset. The training was carried out using 5-fold validation on the training set. Various models built using a combination of hyperparameters for each algorithm were tested on the validation set. Then, the models were tested on a test set to select final

models that would be deployed on a web application built using streamlit library of python. All the machine learning classification algorithms were employed using sci-kit learn 0.22.1 on python 3.7.

10.2.7. Validation of machine learning model

The developed models were validated using accuracy, precision, recall and F1-score as described earlier (Section 1.4.3).

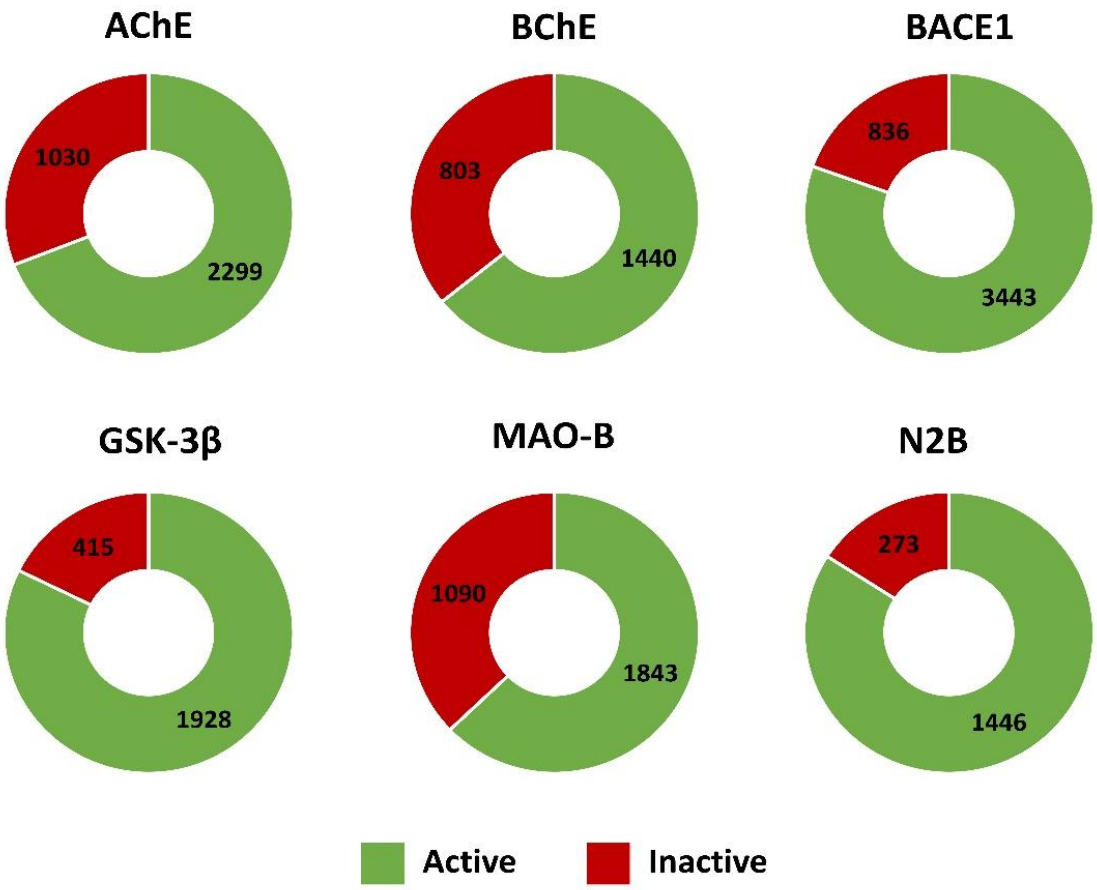

Figure 10.1 Distribution of datasets into active and inactive classes.

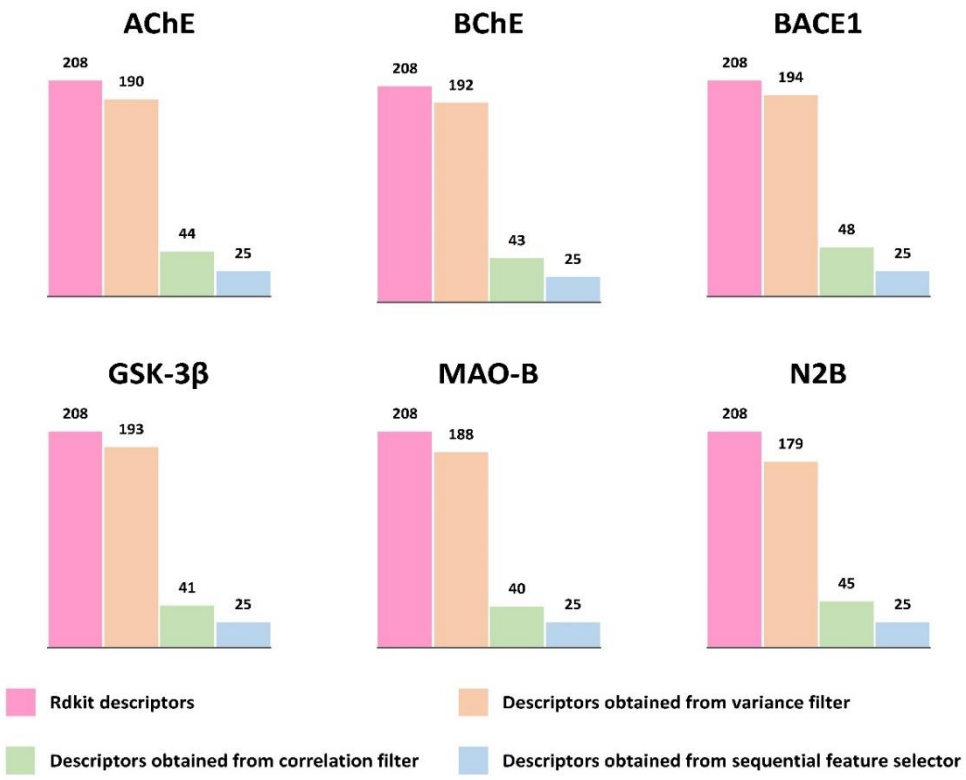

Figure 10.2 Feature selection using filters.

10.3. Results and discussion

10.3.1. Model development for identification of acetylcholinesterase inhibitors

The AChE datasets obtained from the Bindingdb database consisted of 7213 inhibitors. After preprocessing, it resulted in 3329 inhibitors that were divided into active and inactive classes using an IC_{50} cut-off value of 5000 nM as reported in the figure (**Figure 10.1**). Further, the number of RDkit descriptors were reduced using various filters, as indicated in the figure (**Figure 10.2**). The training accuracy of the model showed that KNN, SVC, decision tree and RF classifier performed well on the training dataset with mean accuracies above 80 %. Further evaluation on the validation and test reflected that accuracy remained above 80 %. (**Table 10.2**).

The F1 score was also above 85 % for these models, which indicated that both precision and recall values were high. Finally, the area under the precision-recall curve was above 90% with both datasets, indicating good quality of models. Hence, these four models were selected for the development of the web application.

Table 10.1 RDkit descriptors used for the development of ML models.

Target	Features
AChE	MaxEStateIndex, MinEStateIndex, MinAbsEStateIndex, qed, MolWt, NumRadicalElectrons, MaxPartialCharge, MinPartialCharge, BCUT2D_MWHI, BCUT2D_CHGHI, BCUT2D_CHGLO, BCUT2D_LOGPHI, BalabanJ, HallKierAlpha, Ipc, PEOE_VSA11, PEOE_VSA14, PEOE_VSA2, PEOE_VSA3, PEOE_VSA5, PEOE_VSA7, SMR_VSA4, SlogP_VSA1, SlogP_VSA8, VSA_EState1
BChE	MaxEStateIndex, MinEStateIndex, MinAbsEStateIndex, qed, MolWt, MaxPartialCharge, FpDensityMorgan1, BCUT2D_MWHI, BCUT2D_CHGLO, BCUT2D_LOGPHI, BCUT2D_LOGPLOW, BCUT2D_MRLOW, HallKierAlpha, PEOE_VSA1, PEOE_VSA10, PEOE_VSA11, PEOE_VSA13, PEOE_VSA14, PEOE_VSA5, PEOE_VSA8, SMR_VSA9, SlogP_VSA10, SlogP_VSA11, EState_VSA11, EState_VSA6
BACE1	MinEStateIndex, MinAbsEStateIndex, qed, NumRadicalElectrons, BCUT2D_MWLOW, BCUT2D_CHGHI, BCUT2D_LOGPHI, BCUT2D_LOGPLOW, BCUT2D_MRLOW, HallKierAlpha, PEOE_VSA1, PEOE_VSA13, PEOE_VSA14, PEOE_VSA7, SMR_VSA5, SMR_VSA6, EState_VSA3, EState_VSA6, VSA_EState2, VSA_EState4, VSA_EState9, NumAliphaticHeterocycles, NumAliphaticRings, fr_alkyl_halide, fr_guanido
GSK3B	MinEStateIndex, qed, MolWt, MaxAbsPartialCharge, BCUT2D_MWLOW, BCUT2D_CHGHI, BCUT2D_CHGLO, BCUT2D_LOGPHI, BCUT2D_LOGPLOW, BCUT2D_MRHI, BertzCT, HallKierAlpha, Ipc, PEOE_VSA11, PEOE_VSA13, PEOE_VSA2, PEOE_VSA6, PEOE_VSA7, PEOE_VSA9, SlogP_VSA11, SlogP_VSA2, VSA_EState2, VSA_EState3, NumSaturatedRings, fr_NH1
MAOB	MolWt, MinPartialCharge, FpDensityMorgan1, BCUT2D_MWLOW, BCUT2D_LOGPHI, BCUT2D_LOGPLOW, BCUT2D_MRHI, PEOE_VSA10, PEOE_VSA11, PEOE_VSA2, PEOE_VSA7, SMR_VSA10, SMR_VSA7, SlogP_VSA3, SlogP_VSA4, SlogP_VSA5, EState_VSA8, VSA_EState6, FractionCSP3, NumAromaticHeterocycles, fr_Al_COO, fr_COO, fr_HOCCN, fr_NH1, fr_priamide
N2B	MaxEStateIndex, MinEStateIndex, MinAbsEStateIndex, qed, MolWt, MaxPartialCharge, MinPartialCharge, MaxAbsPartialCharge, FpDensityMorgan1, BCUT2D_MWHI, BCUT2D_CHGHI, BCUT2D_CHGLO, BCUT2D_LOGPHI, BCUT2D_LOGPLOW, BCUT2D_MRHI, BCUT2D_MRLOW, BalabanJ, BertzCT, HallKierAlpha, PEOE_VSA11, PEOE_VSA3, PEOE_VSA4, PEOE_VSA7, SMR_VSA1, VSA_EState6

10.3.2. Model development for identification of butyrylcholinesterase inhibitors

A dataset of 9930 human BChE inhibitors was collected from the Bindingdb database was processed to obtained 2243 compounds. The 25 rdkit descriptors of the compounds that were selected for training are available in the table (**Table 10.1**). The results indicated

that the LR and Naïve Bayes classifier poorly classified the datasets. The low precision and recall indicated that the models could not identify the active compound in the validation and test sets (**Table 10.3**). The other models based on KNN, SVC, decision tree and RF classifier showed accuracy above 80 % on the training set. However, the validation and test set accuracy remained below 80 % for these models except the model developed from the RF classifier. The area under the precision-recall curve was more than 90% for only the RF classifier. Hence, it was the only model selected for deployment on the web application.

10.3.3. Model development for identification of β-secretase 1 inhibitors

The development of the ML model for BACE1 inhibitor's prediction was carried out with a dataset of 4179 BACE1 inhibitors obtained from the Bindingdb database after processing 13481 inhibitors. It was observed that the KNN, SVC, decision tree and RF classifier models performed well in terms of accuracies among all three datasets. The precision and recall values of these models were quite high, i.e., above 90 % on the validation set as well as test set (**Table 10.4**).

The remaining algorithms, viz. LR and Gaussian Naïve Bayes algorithms, did not perform well. However, their precisions were relatively high, but they could not predict the negative sample, i.e., inactive compounds, correctly at a large scale. Further, a single decision tree displayed lower accuracy than the RF classifier that used a combination of various decision trees trained on the subsets of the data. The KNN, SVC, decision tree and RF classifier models were used for a web application.

10.3.4. Model development for identification of glycogen synthase kinase 3β inhibitors

A dataset of 8322 GSK-3β inhibitors was collected and processed to obtain 2343 compounds with their IC_{50} values. Except for the LR, the other models performed well on the training dataset. Further, the F1 score and area under the precision-recall curve

were above 90 %. Hence, the model based on KNN, SVC, decision tree and RF classifier algorithms were selected (**Table 10.5**).

10.3.5. Model development for identification of monoamine oxidase B inhibitors

The ML models for MAO-B were developed from 2933 inhibitors that were obtained after processing the 8176 inhibitors obtained from Bindingdb. The training of the models provided an interesting observation that all the ML models, except the RF classifier, showed training accuracy below 80 %. However, these models showed test and validation accuracies above 80% (**Table 10.6**). Hence, the RF classifier was only selected for the web application. Its F1 and area under the precision-recall curve scores were relatively high.

10.3.6. Model development for identification of N_2B inhibitors of NMDA receptor

A dataset of 3198 N_2B subunit inhibitors was collected and processed to obtain 1719 compounds with their IC_{50} values. The training of the model resulted in all the algorithms showing training accuracies above 80 %. Further, the validation and test sets also reflected similar results (**Table 10.7**). Hence, the model based on KNN, SVC, decision tree and RF classifier algorithms were selected due to their F1 scores above 90 %.

10.3.7. Alzleads

The selected models were deployed in the form of a web application which was built and hosted with the help of Streamlit, a python library. Alzleads is a web application made available to the user through the website (https://www.drugdesign.in/tools/alzleads).

Table 10.2 Performance of the models trained on AChE inhibitor dataset.

		Training set	Validation set					Test set				
Algorithm	**Parameters**	**Accuracy**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**
Logistic regression	C = 29.763, class weight = balanced, solver = liblinear	71.05 ± 1.65	72.94	83.92	74.31	78.82	86.79	70.6	82.88	71.39	76.7	86.79
KNN	algorithm = brute, leaf size =1, metric = euclidean, n_neighbors = 3	83.53 ± 2.05	84	84.59	93.4	88.78	93.28	83.6	84.82	92.33	88.42	93.28
SVC	C = 20, class weight = balanced, gamma = 0.1, probability = True	84.24 ± 1.42	84	87.16	89.58	88.36	90.93	85	88.6	89.38	88.99	90.93
Gaussian Naïve Bayes	-	69.67 ± 0.09	67.76	67.76	100	80.79	76.8	68	67.94	100	80.91	76.8
Decision tree	class weight = balanced, criterion = entropy, max_features = 0.5	82.36 ± 1.18	82.12	86.05	87.85	86.94	91.15	83.4	85.96	90.27	88.06	91.15
RF classifier	class weight = balanced, criterion = entropy, max_features = 0.7, n_estimators = 50	86.86 ± 1.43	86.82	87.18	94.44	90.67	95.73	87.6	87.95	94.69	91.19	95.73

The training set accuracy is represented as Mean ± SD that was obtained from 5-fold validation. PR AUC represents the area under the precision-recall curve.

Table 10.3 Performance of the models trained on BChE inhibitor dataset.

		Training set	Validation set					Test set				
Algorithm	**Parameters**	**Accuracy**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**
Logistic regression	C = 0.61, class weight = balanced, solver = liblinear	67.84 ± 2.07	67.13	73.29	66.05	69.48	75.61	64.39	78.12	65.79	71.43	78.15
KNN	algorithm = ball_tree, leaf size = 4, metric = euclidean, n_neighbors = 3	80.99 ± 0.25	79.02	78.65	86.42	82.35	88.88	79.23	84.05	85.53	84.78	90.87
SVC	C = 10, class weight = balanced, probability = True	81.78 ± 1.73	77.62	79.52	81.48	80.49	85.44	78.34	85.71	81.58	83.6	87
Gaussian Naïve Bayes	-	41.67 ± 2.84	46.5	73.68	8.64	15.47	67.76	37.98	80.65	10.96	19.31	78.55
Decision tree	class weight = balanced, max_features = 0.5, splitter = random	78.14 ± 2.37	78.32	77.47	87.04	81.98	85.95	78.04	83.48	84.21	83.84	89.19
RF classifier	class weight = balanced, max_features = 0.1	83.09 ± 1.87	80.42	78.8	89.51	83.82	90.89	81.01	83.88	89.04	86.38	90.94

The training set accuracy was represented as Mean ± SD that was obtained from 5-fold validation. PR AUC represents the area under the precision-recall curve.

Table 10.4 Performance of the models trained on BACE1 inhibitor dataset.

		Training set	Validation set					Test set				
Algorithm	Parameters	Accuracy	Accuracy	Precision	Recall	F1 score	PR AUC	Accuracy	Precision	Recall	F1 score	PR AUC
Logistic regression	C = 0.033, class weight = balanced, penalty = l1, solver = liblinear	75.22 ± 2.48	74.91	91.85	75.17	82.68	93.8	77.26	92.1	78.61	84.82	93.76
KNN	algorithm = ball_tree, leaf size = 10, metric = euclidean, n_neighbors = 3	86.67 ± 0.77	88.1	89.7	96.09	92.79	95.41	87.69	90.89	94.22	92.53	96.07
SVC	C = 1, class weight = balanced, gamma = 0.1, probability = True	85.02 ± 1.08	86.26	91.86	90.8	91.33	97.36	87.23	94.14	89.79	91.91	97.51
Gaussian Naïve Bayes	-	71.53 ± 12.3	77.84	87.74	83.91	85.78	92.53	78.35	87.7	85.16	86.41	90.49
Decision tree	class weight = balanced, criterion = entropy, max_features = 0.7, splitter = random	82.89 ± 1.2	85.53	91.4	90.34	90.87	94.76	84.11	90.64	89.6	90.12	94.42
RF classifier	class weight = balanced, criterion = entropy, max_features = 0.5	87.48 ± 0.75	89.38	90.36	97.01	93.57	97.81	87.54	89.84	95.38	92.52	97.52

The training set accuracy was represented as Mean ± SD that was obtained from 5-fold validation. PR AUC represents the area under the precision-recall curve.

Table 10.5 Performance of the models trained on GSK-3β inhibitor dataset.

		Training set	Validation set					Test set				
Algorithm	**Parameters**	**Accuracy**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**
Logistic regression	C = 1.62, class weight = balanced, penalty = l1, solver = liblinear	69.67 ± 1.56	67.89	90.58	68.92	78.28	93.08	71.02	87.66	73.84	80.16	90.73
KNN	algorithm = ball_tree, leaf size = 1, metric = euclidean, n_neighbors = 8	85.99 ± 0.44	85.95	88	96.41	92.02	96.89	83.81	84.47	97.49	90.52	95.34
SVC	C = 10, class weight = balanced, gamma = 0.1, probability = True	86.05 ± 1.44	88.96	91.92	95.22	93.54	97.48	86.93	90.03	93.91	91.93	94.66
Decision tree	class weight = balanced, max_features = 0.1	83.1 ± 1.19	82.27	87.79	91.63	89.67	93.22	83.24	87.67	91.76	89.67	92.98
RF classifier	class weight = balanced, max_features = 0.7, n_estimators = 50	87.22 ± 1.63	84.95	87.59	95.62	91.43	96.97	85.51	86.31	97.13	91.4	95.48

The training set accuracy was represented as Mean ± SD that was obtained from 5-fold validation. PR AUC represents the area under the precision-recall curve

Table 10.6 Performance of the models trained on MAO-B inhibitor dataset.

		Training set	Validation set					Test set				
Algorithm	**Parameters**	**Accuracy**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**
Logistic regression	C = 0.03, class weight = balanced, solver = liblinear	64.42 ± 1.76	72.94	83.92	74.31	78.82	86.79	70.6	82.88	71.39	76.7	86.79
KNN	algorithm = ball_tree, leaf size = 1, metric = euclidean	79 ± 1.72	84	84.59	93.4	88.78	93.28	83.6	84.82	92.33	88.42	93.28
SVC	C = 20, class weight = balanced, gamma = auto, probability = True	79.42 ± 1.28	84	87.16	89.58	88.36	90.93	85	88.6	89.38	88.99	90.93
Gaussian Naïve Bayes	-	65.64 ± 5.53	67.76	67.76	100	80.79	76.8	68	67.94	100	80.91	76.8
Decision tree	class weight = balanced, criterion = entropy, max_features = 0.5	74.71 ± 2.07	82.12	86.05	87.85	86.94	91.15	83.4	85.96	90.27	88.06	91.15
RF classifier	class weight = balanced, criterion = entropy, max_features = 0.7, n_estimators = 20	80.89 ± 1.04	86.82	87.18	94.44	90.67	95.73	87.6	87.95	94.69	91.19	95.73

The training set accuracy was represented as Mean ± SD that was obtained from 5-fold validation. PR AUC represents the area under the precision-recall curve

Table 10.7 Performance of the models trained on N2B subunit inhibitor dataset.

		Training set	Validation set					Test set				
Algorithm	**Parameters**	**Accuracy**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**PR AUC**
Logistic regression	C = 0.61, class weight = balanced, solver = liblinear	85.17 ± 1.39	88.64	96.59	89.95	93.15	98.86	81.4	94.05	82.46	87.88	93.76
KNN	algorithm = ball_tree, leaf size = 1, metric = euclidean, n_neighbors = 3	93.07 ± 1.46	94.55	94.92	98.94	96.89	97.58	90.7	91.93	97.16	94.47	96.26
SVC	C = 20, class weight = balanced, gamma = 0.1, probability = True	91.53 ± 2.17	91.82	96.22	94.18	95.19	97.67	90.7	93.9	94.79	94.34	94.49
Gaussian Naïve Bayes	-	86.3 ± 0.86	89.55	95.6	92.06	93.8	94.28	84.88	93.88	87.2	90.42	95.5
Decision tree	class weight = balanced, criterion = entropy, max_features = 0.5	92.91 ± 1.71	92.73	96.26	95.24	95.74	97.79	89.53	93.81	93.36	93.59	96.3
RF classifier	class weight = balanced, max_features = 0.1, n_estimators = 50	94.6 ± 1.53	95.45	95.9	98.94	97.4	98.95	92.64	93.24	98.1	95.61	95.91

The training set accuracy was represented as Mean ± SD that was obtained from 5-fold validation. PR AUC represents the area under the precision-recall curve