CS4742 - Bioinformatics

Prediction of Druggable Proteins

Report

Group Celestials

180070L - L.C.I. Bannaheke
180118T - D.Y. de Sliva
180273L - T.T. Jayasekara
180449H - K.P.D.T. Pathirana

Table of Content

Q1

A ***Druggable Protein*** is a protein that can interact or attach with drug-like molecules and can result in a desired state in medicinal treatments. Therefore, identifying druggable proteins is a huge asset in the drug industry. But just plainly identifying them using traditional experiments is costly and time-consuming. Therefore, it is proposed to determine different features of druggable proteins that can help identify them and then, develop machine learning models using these features to predict the druggability of a given protein sequence. These features can be extracted using different properties of proteins such as physicochemical properties, compositional information, and composition-transition-distribution information. Some features that have been identified to outline the nature of druggable proteins are mentioned below.

Feature	Description	Dimensions
AAC [1]	*Amino Acid Composition* - Calculates the frequency of each type of amino acid.	20
PAAC [1]	*Pseudo Amino Acid Composition* - Converts protein sequences into fixed-length numerical vectors by incorporating the physicochemical properties of amino acids and their sequence order information.	21
APAAC [1]	*Amphiphilic Pseudo-Amino Acid Composition* - Combines sequence order and physicochemical properties, using descriptors based on the presence and weighted properties of amino acids in overlapping segments, to capture the amphiphilic nature of proteins.	22
CTD [1]	*Composition, Transition,* and *Distribution* - The Composition descriptor focuses on the overall proportion of a specific amino acid attribute group, the Transition descriptor measures the frequency of transitions between different attribute groups, and the Distribution descriptor examines the distribution patterns of a particular attribute group within the protein sequence.	273
DPC [1]	*Dipeptides Composition* - Calculated using the percentages of the 400 dipeptide combinations.	400
TPC [1]	*Tripeptide Composition* - Reflects the statistical frequency of any combination of three amino acids.	8000
GAAC [2]	*Grouped Amino Acid Composition*	5
GDPC [3]	*Grouped Dipeptide Composition* - Groups into five classes using their physicochemical properties as follows: Aromatic, Positive Charge, Aliphatic, Uncharged, and Negative Charged.	5
RAAA [3]	*Reduced Amino Acid Alphabet* - By utilizing the physiochemical properties, the amino acid residues were categorized into smaller groups. This categorization not only decreased the complexity of protein sequences but also facilitated the exploration of structural local regions and identified structural similarities.	5

RSacid [1]	*Reduced* amino acid *Sequences* according to *acidity.*	32
RScharge [1]	*Reduced* amino acid *Sequences* according to *charge.*	50
RSDHP [1]	*Reduced* amino acid *Sequences* according to *DHP.*	32
RSpolar [1]	*Reduced* amino acid *Sequences* according to *polarity.*	32
RSsecond [1]	*Reduced* amino acid *Sequences* according to *secondary structure.*	40
monoDiKGap [2]	A variant of the Kmer feature extraction method.	16000

Q2

Used Features

A set of 10 features (from the above list in Q1) were selected to be extracted and used individually with different classification models in order to predict the druggability of proteins. The selected features are as follows.

1. AAC
2. APAAC
3. CTD
4. DPC
5. PAAC
6. RSacid
7. RScharge
8. RSDHP
9. RSpolar
10. RSsecond

Used Classification Models

For each selected feature, seven different classification models were developed and all of them were trained and validated using the given dataset. Used classification models are as follows.

- **Extra Trees** [4]
- **KNN**: K - Nearest Neighbours [5]
- **LightGBM**: Light Gradient Boosting Machine [6]
- **LR**: Linear Regression [7]
- **RF**: Random Forest [8]
- **SVC**: Support Vector Classification [9]
- **XGBoost**: Extreme Gradient Boosting [10]

Test Results for Each Feature

After training and validating the developed models, several different evaluation metrics were obtained as mentioned in the below table. The highlighted models are the best-performed model for each of the features according to their F1 Score.

Feature	Classification Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
AAC	Extra Trees	0.88277	0.86754	0.89689	0.88297	0.88250
	LightGBM	0.88316	0.87408	0.89158	0.88311	0.88296
	RF	0.88080	0.85037	0.90902	0.88207	0.88033
	LR	0.84579	0.87163	0.82183	0.84643	0.84578

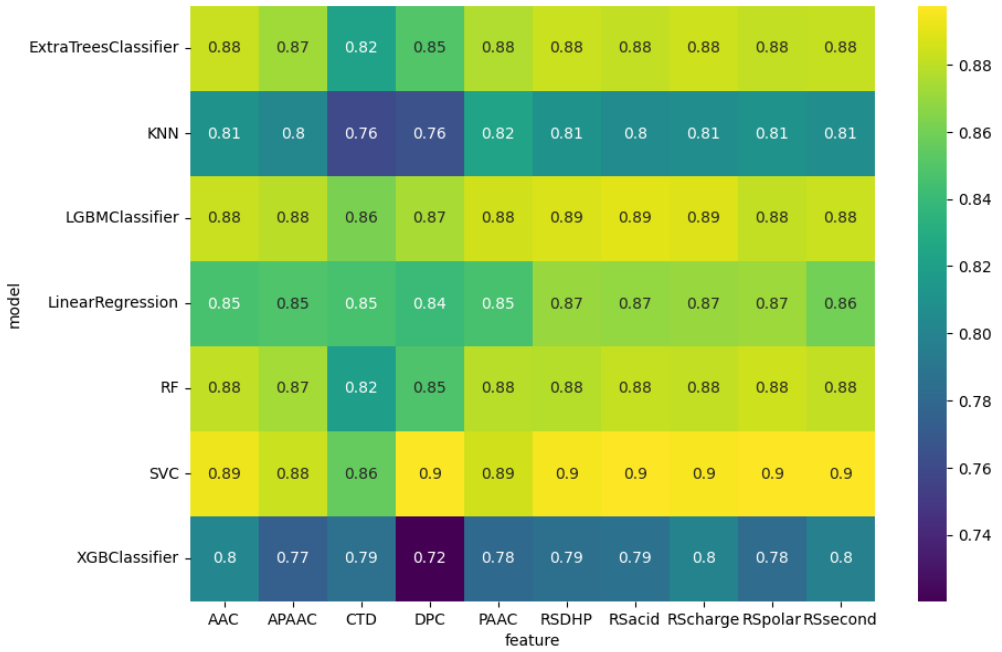
	XGBoost	0.80212	0.78741	0.81577	0.80195	0.80173
	SVC	0.89339	0.87980	0.90599	0.89356	0.89316
	KNN	0.81078	0.78168	0.83776	0.81124	0.81013
APAAC	Extra Trees	0.87215	0.86345	0.88021	0.87205	0.87193
	LightGBM	0.87962	0.87163	0.88704	0.87953	0.87942
	RF	0.87411	0.85854	0.88855	0.87428	0.87382
	LR	0.85090	0.75552	0.93935	0.86296	0.84858
	XGBoost	0.77341	0.80376	0.74526	0.77451	0.77341
	SVC	0.88395	0.85037	0.91509	0.88556	0.88345
	KNN	0.80566	0.69910	0.90447	0.81791	0.80217
CTD	Extra Trees	0.82297	0.80703	0.83776	0.82291	0.82259
	LightGBM	0.86192	0.84219	0.88021	0.86223	0.86155
	RF	0.81983	0.79150	0.84610	0.82032	0.81921
	LR	0.84579	0.89616	0.79909	0.84888	0.84575
	XGBoost	0.78600	0.81848	0.75588	0.78725	0.78599
	SVC	0.85602	0.84710	0.86429	0.85588	0.85578
	KNN	0.76042	0.78005	0.74223	0.76085	0.76040
DPC	Extra Trees	0.84972	0.85609	0.84382	0.84953	0.84962
	LightGBM	0.87490	0.86182	0.88704	0.87496	0.87464
	RF	0.84815	0.82829	0.86657	0.84838	0.84775
	LR	0.84068	0.87899	0.80516	0.84237	0.84067
	XGBoost	0.72030	0.73671	0.70508	0.72063	0.72026
	SVC	0.89693	0.87980	0.91281	0.89732	0.89667
	KNN	0.76593	0.71872	0.80970	0.76713	0.76463
	Extra Trees	0.87648	0.86345	0.88855	0.87654	0.87622
	LightGBM	0.88474	0.87572	0.89310	0.88469	0.88453

PAAC	RF	0.87884	0.86509	0.89158	0.87895	0.87857
	LR	0.84736	0.77187	0.91736	0.85456	0.84567
	XGBoost	0.78049	0.78496	0.77635	0.78029	0.78035
	SVC	0.88552	0.85119	0.91736	0.88724	0.88501
	KNN	0.82612	0.72608	0.91888	0.83796	0.82325
RSacid	Extra Trees	0.88120	0.86590	0.89538	0.88139	0.88092
	LightGBM	0.88946	0.87653	0.90144	0.88958	0.88922
	RF	0.88238	0.85037	0.91205	0.88381	0.88189
	LR	0.86861	0.89125	0.84761	0.86900	0.86859
	XGBoost	0.78757	0.79068	0.78469	0.78734	0.78742
	SVC	0.89732	0.87572	0.91736	0.89801	0.89702
	KNN	0.80566	0.77351	0.83548	0.80627	0.80493
RScharge	Extra Trees	0.88395	0.86836	0.89841	0.88417	0.88368
	LightGBM	0.88867	0.87326	0.90296	0.88891	0.88841
	RF	0.88198	0.85119	0.91054	0.88330	0.88151
	LR	0.87018	0.89861	0.84382	0.87095	0.87018
	XGBoost	0.79976	0.85282	0.75057	0.80318	0.79968
	SVC	0.89575	0.87735	0.91281	0.89621	0.89547
	KNN	0.80685	0.77923	0.83245	0.80719	0.80621
RSDHP	Extra Trees	0.88395	0.86509	0.90144	0.88436	0.88364
	LightGBM	0.88749	0.87490	0.89917	0.88759	0.88726
	RF	0.87844	0.84383	0.91054	0.88009	0.87790
	LR	0.87018	0.89534	0.84685	0.87072	0.87017
	XGBoost	0.78521	0.79477	0.77635	0.78515	0.78512
	SVC	0.89536	0.87326	0.91585	0.89607	0.89504
	KNN	0.81078	0.80049	0.82032	0.81054	0.81047

RSpolar	Extra Trees	0.88120	0.86590	0.89538	0.88139	0.88092
	LightGBM	0.88159	0.86999	0.89234	0.88162	0.88135
	RF	0.88434	0.85282	0.91357	0.88575	0.88387
	LR	0.87136	0.89452	0.84989	0.87177	0.87135
	XGBoost	0.78245	0.82093	0.74678	0.78424	0.78244
	SVC	0.89772	0.87326	0.92039	0.89863	0.89738
	KNN	0.80999	0.82747	0.79378	0.81022	0.80996
RSsecond	Extra Trees	0.88198	0.86590	0.89689	0.88222	0.88170
	LightGBM	0.88356	0.87326	0.89310	0.88355	0.88334
	RF	0.88198	0.84955	0.91205	0.88345	0.88149
	LR	0.85995	0.89207	0.83017	0.86104	0.85995
	XGBoost	0.79701	0.82584	0.77028	0.79796	0.79701
	SVC	0.89772	0.88471	0.90978	0.89789	0.89750
	KNN	0.80724	0.78741	0.82563	0.80724	0.80676

Accuracy Comparison Heatmap

The following figure illustrates accuracy values obtained for each feature and model. The light color areas indicate high accuracy whereas the dark color indicates a low accuracy. It can be clearly seen that almost all the SVC models indicate a light color suggesting that the SVC classification model performs best for many features.



Best-Performed Models

Altogether 70 (*10 * 7*: 10 features and 7 classification models) machine learning models were developed and out of them, the best-performed model for each feature was identified using their F1 Score.

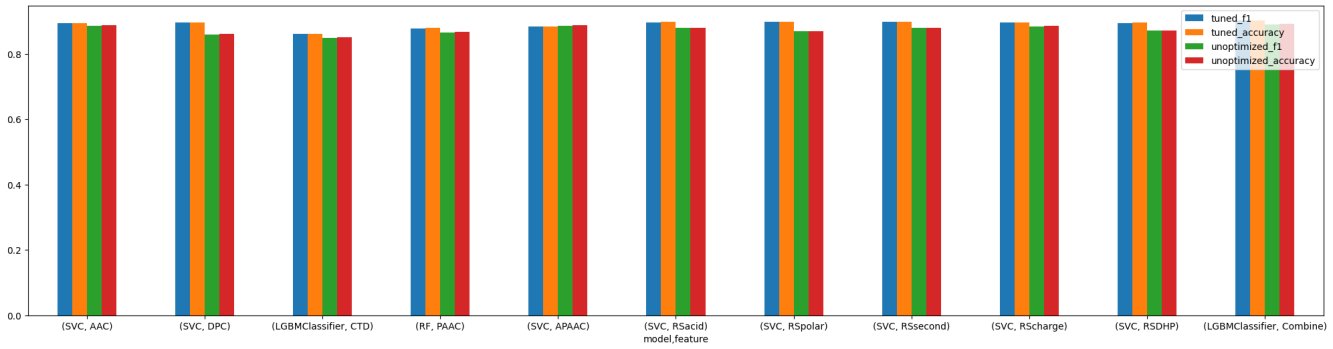
Hyper Parameters

The hyperparameters were tuned for each of the best-performed models and the obtained parameters and values are mentioned below.

Feature	Best Performed Model	Hyper Parameters	
		Parameter	Value
AAC	SVC	C	1.7973097772791014
APAAC	SVC	C	10.048020657539734
CTD	LightGBM	n_estimators	283
		max_depth	54
		num_leaves	22
DPC	SVC	C	5.898246724239805
PAAC	RF	n_estimators	230
		max_depth	49
		criterion	log_loss
RSacid	SVC	C	1.9242211723416096
RScharge	SVC	C	1.9897995542245812
RSDHP	SVC	C	2.8945067227904238
RSpolar	SVC	C	2.9411040169428166
RSsecond	SVC	C	1.4295599735253435

Graphical Representation of Test Results

The following figure is a representation of F1 score and accuracy of all best-performed models before and after tuning. In almost all cases, the tuned F1 score and accuracy are greater than that of the untuned versions.



Best Performed Model of Q2

The final best-performed model out of all the best-performed models in Q2 is the SVC model developed for RSsecond feature and its test results are as follows.

- **Model**: RSsecond-SVC
- **Features**: RSsecond
- **Accuracy**: 0.89772
- **Sensitivity**: 0.88471
- **Specificity**: 0.90978
- **Precision**: 0.89789
- **F1 Score**: 0.89750

Q3

In this phase, a set of models were developed by combining different selected features and it was done in two ways.

1. *Feature Combination at the Level of Features* - Here, the selected features were combined into a single feature vector and a set of classification models were trained for that combined feature.
2. *Feature Combination with Ensemble of Classifiers* - Here, a set of classifiers were trained for each of the selected features and built an ensemble of those classifiers into a single model.

Feature Combination at the Level of Features

Two feature combinations were performed where one considered all 10 features and the other considered only the best-performed five features in Q2. For both cases, seven models were built using each of the classifiers used in Q2. The test results for each feature combination and each classification model is as follows. The highlighted models are the best-performed ones for each feature combination.

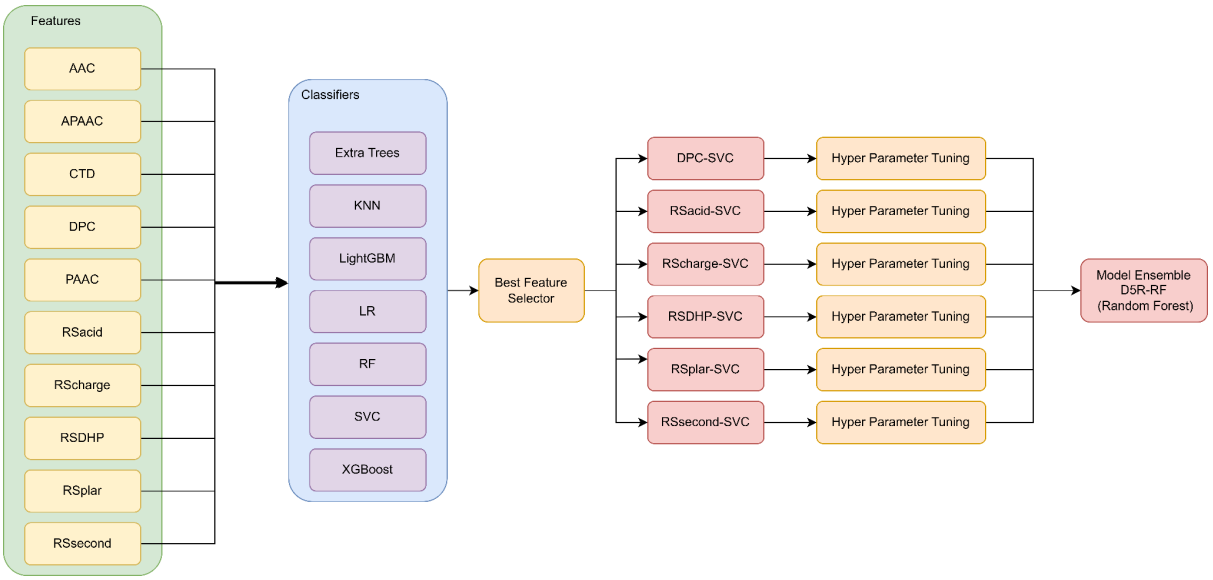
Features	Classification Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
All 10 Features Combined	Extra Trees	0.88041	0.86999	0.89007	0.88039	0.88018
	LightGBM	0.90283	0.88471	0.91964	0.90332	0.90257
	RF	0.88277	0.86182	0.90220	0.88330	0.88243
	LR	0.81943	0.86182	0.78014	0.82158	0.81942
	XGBoost	0.83596	0.83320	0.83851	0.83569	0.83576
	SVC	0.90362	0.88062	0.92494	0.90447	0.90331
	KNN	0.81039	0.83974	0.78317	0.81135	0.81038
DPC + RSDHP + RSacid + RSpolar + RSsecond + RScharge (D5R)	Extra Trees	0.89457	0.88307	0.90523	0.89465	0.89436
	LightGBM	0.81235	0.83892	0.78772	0.81311	0.81235
	RF	0.90165	0.88553	0.91660	0.90201	0.90141
	LR	0.84500	0.87653	0.81577	0.84607	0.84500
	XGBoost	0.89142	0.86345	0.91736	0.89258	0.89102
	SVC	0.90480	0.88716	0.92115	0.90527	0.90455
	KNN	0.80330	0.77105	0.83321	0.80389	0.80256

Feature Combination with Ensemble of Classifiers

A Random Forrest classifier was developed and trained with the following ensemble of classifiers developed and trained in Q2.

- DPC-SVC
- RSDHP-SVC
- RSacid-SVC
- RSpolar-SVC
- RSsecond-SVC
- RScharge-SVC

The D5R-RF Model Architecture



Model	Classification Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
Ensemble	RF	0.83297	0.77232	0.89030	0.83735	0.83183

Best Performed Model of Q3

The final best-performed model out of all the best-performed models in Q3 is the SVC model developed for DPC, RSDHP, RSacid, RSpolar, RSsecond, RScharge features, and its test results are as follows.

- **Model**: D5R-SVC
- **Features**: DPC, RSDHP, RSacid, RSpolar, RSsecond, RScharge
- **Accuracy**: 0.90480
- **Sensitivity**: 0.88716
- **Specificity**: 0.92115
- **Precision**: 0.90527
- **F1 Score**: 0.90455

Q4

The McNemar test [11] was used to compare the best-performed model identified in Q3 (D5R-SVC) with best-performed model identified in Q2 (RSsecond-SVC). The McNemar test is a non-parametric test to assess if there is a statistically significant change in proportions and this method is appropriate to compare the above two models on the same dataset.

In this test, the null hypothesis is formulated as no model performs better than the other model. Therefore, the alternative hypothesis is that the two models perform differently. In conducting the McNemar test, the predictions from RSsecond-SVC model and D5R-SVC model are collected and then a 2x2 contingency table is prepared as shown below.

	D5R-SVC Correct	**D5R-SVC Incorrect**
RSsecond-SVC Correct	373	11
RSsecond-SVC Incorrect	35	42

The table represents that the number of cases where both models predicted correctly is 382, the number of cases where both models predicted wrong is 42, the number of cases where RSsecond-SVC predicted correctly while D5R-SVC predicted incorrectly (f) is 11 and the number of cases D5R-SVC predicted correctly while RSsecond-SVC predicted incorrectly (s) is 35.

The test statistic was calculated using the equation, $\chi^2 = (|s - f| - 1)^2 / (s + f)$ as 12.5217 and the P value was computed assuming that the null hypothesis is true, as 0.0004022443020605948.

We set the significance threshold as 0.01 and the computed P value being lower than the chosen significance level allows us to reject the null hypothesis. Further, the 11:35 ratio allows us to conclude that **the D5R-SVC model performs substantially better than the RSsecond-SVC model**.

REFERENCES

[1] P. Charoenkwan, N. Schaduangrat, P. Lio', M. A. Moni, W. Shoombuatong, and B. Manavalan, "Computational prediction and interpretation of druggable proteins using a stacked ensemble-learning framework," iScience, vol. 25, no. 9, p. 104883, Sep. 2022, doi: https://doi.org/10.1016/j.isci.2022.104883.

[2] Y.-X. Gong, B. Liao, P. Wang, and Q. Zou, "DrugHybrid_BS: Using Hybrid Feature Combined With Bagging-SVM to Predict Potentially Druggable Proteins," vol. 12, Nov. 2021, doi: https://doi.org/10.3389/fphar.2021.771808.

[3] R. Sikander, A. Ghulam, and F. Ali, "XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set," vol. 12, no. 1, Apr. 2022, doi: https://doi.org/10.1038/s41598-022-09484-3.

[4] "3.2.4.3.3. sklearn.ensemble.ExtraTreesClassifier — scikit-learn 0.22.2 documentation," scikit-learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

[5] JavaTpoint, "K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint," www.javatpoint.com, 2021. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

[6] "Welcome to LightGBM's documentation! — LightGBM 3.3.2 documentation," lightgbm.readthedocs.io. https://lightgbm.readthedocs.io/en/v3.3.2/

[7] "sklearn.linear_model.LinearRegression — scikit-learn 0.22 documentation," Scikit-learn.org, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[8] JavaTpoint, "Machine Learning Random Forest Algorithm - Javatpoint," www.javatpoint.com. https://www.javatpoint.com/machine-learning-random-forest-algorithm

[9] scikit-learn developers, "sklearn.svm.SVC — scikit-learn 0.22 documentation," Scikit-learn.org, 2019. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[10] "XGBoost Documentation — xgboost 1.5.1 documentation," xgboost.readthedocs.io. https://xgboost.readthedocs.io/en/stable/ (accessed Jan. 06, 2022).

[11] S. Raschka, "mcnemar: McNemar's test for classifier comparisons - mlxtend," rasbt.github.io. https://rasbt.github.io/mlxtend/user_guide/evaluate/mcnemar/