How to Cite:
Nadia, N., Gandotra, E., & Kumar, N. (2022). Comparative analysis of machine learning based methods for the prediction of NLR protein. *International Journal of Health Sciences*, *6*(S8), 5303–5318. Retrieved from https://sciencescholar.us/journal/index.php/ijhs/article/view/13445

Comparative analysis of machine learning based methods for the prediction of NLR protein

Nadia
Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat- 173234, Solan (HP), India

Ekta Gandotra
Department of Computer Science Engineering &- Information Technology, Jaypee University of Information Technology, Waknaghat- 173234, Solan (HP), India
Correspondence author email: ekta.gandotra@gmail.com

Narendra Kumar
Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat- 173234, Solan (HP), India

***Abstract*---**In intestinal tissue repair and innate immunity, the nucleotide-binding domain leucine-rich repeat-containing (NLR) proteins play a fundamental role. The NLR protein family is a recent addition to the members of innate immunity effector molecules. It also plays an important role in intestinal microbiota, and recently emerged as a crucial hit for the development of colitis-associated cancer (CAC) and ulcerative colitis (UC). We have developed a Machine Learning based method for the prediction of NLR Proteins. This paper presents a comparative analysis of three supervised machine learning algorithms i.e. Sequential Minimal Optimization (SMO), Library for Support Vector Machine (LIBSVM) and Random Forest (RF) for prediction of NLR proteins. The dataset used for this work is created after extracting the features using ProtR package. The models are trained with the input compositional features generated using dipeptide composition, amino acid composition, etc. The dataset employed for training consists of 390 proteins. It has positive (103 sequences) set consisting of sequences from the NLR family and the remaining dataset (287 sequences) act as a negative training set, which has random protein sequences and several transporter family protein sequences retrieved from the NCBI and Uniprot. In the test set, there are 99 protein sequences in all (26 in positive and 73 in the negative set). The five-fold cross-validation (CV) is used to optimize LIBSVM, SMO and Random Forest parameters, and the best model was selected. The proposed NLRPred system performs rationally well

with an accuracy of 90.91% for RF as the best classifier for Amino Acid Composition (AAC) based model. The proposed work suggested a rational and rapid classification of NLR Protein. We believe that NLRPred is reliable, useful and rapid prediction method for NLR Protein.

***Keywords*---**NLR, machine learning, SVM, SMO, random forest, cross-validation.

Introduction

The impact of microbiota on the CAC and intestinal inflammation has increased in the past few years by using the NLR protein (Adachi et al., 2019). The innate immunity system is conserved in animals and plants and acts as the first line of defense against microorganisms (Hirota et al., 2011). A diverse community of microbiomes is made up of a huge number of micro-organisms and bacteria that are used as a host to the human intestine and developed mutually with the intestinal immune system (Baggs, Dagdas, & Krasileva, 2017; Seo et al., 2015). Throughout this mechanism, the immune system of the host and the microbiota form additional support by a different mechanism to get an unbeaten mutual association (Zahid, Li, Kombe, Jin, & Tao, 2019). Dysbiosis is a process that describes the synchronized colonization breakdown, and the evidence suggests that it is the result of inflammatory and contagious disorders (Biswas & Kobayashi, 2013). The cytoplasmic microbial sensors from a diverse family belonging to NLR proteins are implicated in various disorders of mutations together with inflammatory bowel diseases (Chen, 2014). In the other organs and the intestine, micro-organisms and their products are detected by pattern recognition receptors (Nadia & Jayashree, 2020), for instance, NLR proteins and Toll-like receptors to bring out first host defense resp0onses (Liao & Schneider, 2019). The three domains of NLRs, which are characterized by a carboxyl-terminal LRR, fundamental NBD, and an amino-terminal protein action domain(Levy, Stedman, Deutsch, Donnadieu, & Virgin, 2020). NLR family caspase recruitment domain containing protein (NLRCs) and NODs family pyrin domain-containing proteins (NLRPs) with several subfamily members (Liao & Schneider, 2019).The NLR protein has been sub-classified based on their protein-interaction amino-terminal domain such as PYRIN domains (NLRPs) which is found in the NLR protein, caspase-recruitment domains ((CARDs) NLRCs)) contains NLRs (Higashi, Sun, & Ishibashi, 2019), caspase-recruitment domains ((CARDs) NLRCs)) or other NLR family proteins. NLR family proteins holds LRRs for ligand recognition; a NOD domain (also known as NACHT domain); a domain for initiation of signaling, namely pyrin domain; baculovirus inhibitor of apoptosis repeat (BIR) domains or caspase activation and recruitment domain (Kumar, Gromiha, & Raghava, 2008).The NLR family, pyrin domain- containing 3 (NLRP3) is the most largest, studied and best-characterized inflammasome during depression and during inflammation (Xu, Liang, Liao, Chen, & Chang, 2018). However, the experimental determination is labor-and time-extensive as well as requires proper infrastructure. Nowadays, machine learning is an alternative, faster and reliable solution to such problems. Machine learning (ML) deals with the creation and evaluation of algorithms that facilitate pattern recognition,

classification, and prediction based on models derived from the existing data(Fletcher et al., 2019). As is known, ML methods generate a model from the training sample and then subsequently predict the label of the testing sample (Agius, Brieghel, & Andersen, 2020; Ramana & Gupta, 2010a). In this paper, comparative analysis of three machine learning algorithms i.e., SMO, LIBSVM, and Random Forest are carried out for prediction of NLR proteins. The dataset used for this work is created after extracting the features using ProtR package. This study represents the first effort to identify NLR Protein using machine learning. Moreover, a dataset including NLR and Non-NLR protein samples was created in this work. This dataset could also be used for additional NLR prediction studies (Amouri, Alaparthy, & Morgera, 2020). The models are constructed and trained by using various compositional features such as Amino acid composition (AAC), Dipeptide Composition etc. Particularly with the increasing significance and biotechnological application of NLR, we constructed and optimized a prediction model that will be helpful to the research community and specifically for the biological community (Nudel et al., 2021).The incumbent limitations of experimental methods, time, cost, and the coupled with the tremendous biological significance. It increases interest in proteins that have motivated attempts to develop computational methods to predict NLR Protein (Kigka et al., 2019). Through various research work carried out on NLR proteins, the importance of NLR proteins is realized. Some scientists have also described the variation in NLR copy number across plant, families, species, which in turn support the role of tandem duplication in NLR CNV (Kalita et al., 2008).

Materials and Methods

This section discusses the methodology used for the prediction of NLR protein.

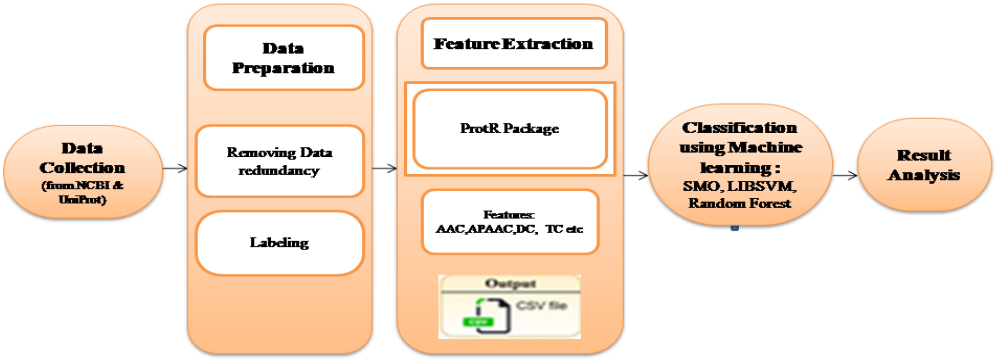

Fig 1: Methodology used for the prediction of NLR Protein

After collecting the data from NCBI and Uniprot, the redundancy is removed and features are extracted with the help of ProtR Package. The whole dataset is split into training and testing dataset. A 5-fold cross validation is used while classifying the data using different ML algorithms i.e., SMO, LIBSVM and Random Forest. The detailed steps of used methodology are given below:

Data Collection

Two sets containing NLR and non-NLR sequences are compiled from National Centre for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov) and Uniprot. The positive set is equipped through keyword search like 'NLR protein' while the negative set is assembled by arbitrarily picking the non-NLR protein. Both the sets are manually inspected to avoid the mislabeling of the data

Data Preparation

Raw pools of sequences for NLR protein are compiled from the database of NCBI with the help of keyword search and literature survey. The entries of filtered sequences were annotated as 'partial', 'hypothetical', 'truncated', 'putative', 'similar to', 'fragment', etc., and after that, the sequences belonging to these criteria are removed from the dataset. Using CD-HIT (Kourou, Exarchos, Exarchos, Karamouzis, & Fotiadis, 2015), the redundancy of the sequences from both sets is scaled to a 50% threshold, yielding 129 NLR (positive set) and 360 Non-NLR (negative set). CD-HIT is a tool for clustering and comparing huge biological sequencing datasets that is frequently used.

Feature Extraction

Followed by the data preparation step, features are extracted using ProtR (R script) package. 11 features namely Amino Acid Composition (AAC), Amphiphilic Pseudo Amino Acid Composition (APAAC), CTDComposition (CTDC), CTDDistribution (CTDD), Conjoint Triad Descriptor (Ctriad), CTDTransition(CTDT), Dipeptide Composition (DC), Geary, Moran, Normalized Moreau-Bruto autocorrelation (Moreau Bruto) and Tripeptide Composition (TC) are extracted with the help of the ProtR Package.

Serial No.	Name of Feature	Dimensions (Amount of Attributes)
1	AAC: Amino acid composition	20
2	APAAC: Amphiphilic Pseudo-amino acid	80
3	C/T/D Composition	39
4	C/T/D Distribution	195
5	C/T/D Transition	39
6	CTriad: Conjoint triad	343
7	DC: Dipeptide composition	400
8	Geary Autocorrelation	240
9	Moran Autocorrelation	240
10	NMBroto: Normalized Moreau Bruto	240
11	TC: Tripeptide composition	8000
		9836

Table1: A brief summary of the selected 11 features retrived from ProtR package.

Classification using Machine Learning

ML has become a significant tool to interpret large and complex biological datasets. Because of this, it has become the go-to tool (Hartmann & Baumert, 2019). It has been used to compact with a wide array of bioinformatics applications. As time has progressed, the profusion of the biological dataset has increased, which has led to the need for novel approaches, as opposed to classical conventional methods (JM & Boylan, 2019). ML is the ability of computers to "learn" from "data" or "experience". It is a collection of models, methods, and algorithms to help make better decisions that are driven by data(Ramana & Gupta, 2010b). As is known, ML methods can learn a model from a training sample and then subsequently predict the label of the testing samples (Jagga & Gupta, 2015). We have pre-processed our data by using Waikato Environment for Knowledge Analysis (WEKA) version 3.6.10, and also used this tool for feature selection and classification procedures. It is an admired suite for ML algorithms that has different tools for evaluation, data pre-processing, association rules, clustering, and visualization (Molteni et al., 2019). We have applied machine learning algorithms like SMO, LIBSVM, and Random Forest.

SMO

This classification algorithm was coined by John C. Platt in 1998 and has become the fastest Quadratic Programming (QP) optimization algorithm, especially for sparse data performance and linear SVM. The SMO algorithm is derived by taking the idea of the decomposition method to its extreme and optimizing a minimal subset of just two points at each iteration. It is widely used for training SVM and is implemented by the popular LIBSVM tool (Tamanna & Ramana, 2015). Now-a-days, the computer programs that provide the best prediction performance are support vector machines (SVMs). This is because SVMs are introduced to maximize the margin to divide two classes so that the trained model generalizes well on unseen data (Agius et al., 2020).

LIBSVM

It is a library of a Support Vector Machine. It is currently one of the most widely used SVM software. A typical use of LibSVM involves two steps: first training a dataset to obtain a model and second, using the model to predict information of testing dataset. It implements the SMO algorithms for kernalized SVM supporting classification and regression (Jagga & Gupta, 2015). LibSVM is a wrapper class for the LibSVM library that supports the classifiers implemented in the LIBSVM library, including one-class SVMs. LibSVM runs faster than SMO. LibSVM allows users to experiment with Regressing SVM, one-class SVM, and nu-SVM supported by LibSVM tool.

SMO and LIBSVM allow us to select a number of kernels and parameters. (e.g. Polynomial, linear, radial basis function (RBF), sigmoid or any user-defined kernel). In this work, we have used RBF kernel. The best consignment is achieved by the optimization of the RBF kernel by altering its parameters (C and gamma) (Tamanna & Ramana, 2015). Here C is a regularization parameter that pedals the trade-off between maximization of the margin and minimization of the training

error and the gamma parameters can be seen as the converse of the radius of influence of samples selected by the model as support vectors (Jagga & Gupta, 2014). RBF is used extensively because of its efficiency and effectiveness (Ramana & Gupta, 2009). We endeavor to select the best model with the best parameters to achieve this or maximize the accuracy and get almost equal specificity and sensitivity wherever possible.

Random Forest

It creates a set of decision trees from the randomly selected subsets. It is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees (JM & Boylan, 2019).The term came from random decision forests that was first proposed by Tin Kam Ho of Bell Labs in 1995.It can be used for classification and regression. It is one of the most accurate learning algorithms available (Ramana & Gupta, 2009). For many data sets, it produces a highly accurate classifier. It runs efficiently on large databases (Ramana & Gupta, 2010a).It can handle thousands of input variables without variable deletion. It also gives estimation of the variables which are significant in the Classification. It generates an internal unbiased estimate of the generalization error as the forest building progresses. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data is missing. RF is fast to build even faster to predict (Tiwari et al., 2020).

Evaluation parameters

A ProtR package was used for feature calculation and WEKA is used for analysis on testing and training set. The complete dataset of the NLR protein sequence is classified into two sets i.e., training and test set. For the optimization of different SVM parameters, a 5-fold Cross Validation method is used. The performance parameters used in this work are True Positive Rate (TPR), False Positive Rate (FPR), Precision, F-Measure, Mathew's Correlation Coefficient (MCC) and Area under Curve (AUC). Their brief description is as follows:

- TPR: It is the ratio of positive instances that are predicted correctly. The number of NLR protein sequences that are predicted correctly as NLR. It is also known as sensitivity or recall.

$$TPR = \frac{TP}{TP + FN}$$

Here, TP is the number of NLR protein sequences that are predicted correctly as NLR and FN is the number of NLR protein sequences that are predicted wrongly as NLR protein.

- FPR: It is the ratio of the number of negative events wrongly predicted as positive and the total number of actual negative events

$$FPR = \frac{FP}{FP + TN}$$

Here, FP is the number of non-NLR protein sequences that are predicted incorrectly as NLR and TN is the number of non-NLR protein sequences that are predicted correctly as non-NLR protein.

- Accuracy (%): It is the percentage of correct predictions for NLR as well as non- NLR sequences.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100$$

- Precision: It is the fraction of relevant instancesamong the retrieved instances. It is also called positive predictive value.

$$Precision = \frac{TP}{TP + FP}$$

- F-measure: It provides a way to combine both precision and recall into a single measure.

$$F - measure = \frac{precision \times recall}{precision + recall} \times 2$$

- MCC: It is employed for the optimization of parameters and evaluation of performance.

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}$$

Here, MCC=1 signifies the perfect prediction while MCC=0 suggests completely random prediction.

- Area under Curve (AUC): The AUC is the measure of the capability of a classifier to differentiate between classes and is used as a summary of the ROC curve (plot between TPR and FPR). The higher the AUC, the better the performance of the model at distinguishing between the positive and negative class.

Results

In this study, we propose the foremost report of ML based method for the identification of NLR protein sequences. Our method minimally represents a complementary method to allow the prediction of NLR protein. For carrying out the classification process, the dataset is splited into two parts i.e., training and testing set. 80% sequences are used for constructing the training set and 20% sequences for the test set. The non-NLR proteins are also compiled using an analogous process. 99 protein sequences are there in the test set in all (26 in the positive and 73 in the negative class) and 390 protein sequences in the training

set (103 in the positive and 287 in the negative class). Though, the ML based methods do come at the cost of some false positive predictions, which should be as minimal as possible. The classification results of the datasets created using ProtR Package package was analyzed using the evaluation parameters TPR, FPR, accuracy, precision, MCC, F-measure, and AUC. After analyzing the results of the training data, the best feature is selected and is used for the classification of the testing data.

Classification Results for training data

This sub-section presents the classification results of the dataset created using features obtained with the help of ProtR Package for training data. Three machine learning classifiers i.e., SMO, LIBSVM, and RF are used for the classification purpose.

Table1 shows the Classification results for SMO on training dataset. The highest accuracy is achieved by the optimization of the classifier with the RBF kernel for DC feature which is described as 93.08% accuracy for $\gamma = 0.09$ and $C = 250$.

Feature	Accuracy (%)	TPR	FPR	Precision	F-Measure	MCC	AUC
AAC	91.03	0.91	0.144	0.91	0.91	0.768	0.883
APAAC	91.54	0.915	0.142	0.915	0.915	0.78	0.886
CTDC	90.51	0.905	0.152	0.905	0.905	0.755	0.876
CTDD	86.67	0.867	0.259	0.863	0.863	0.643	0.804
Ctriad	92.56	0.926	0.157	0.925	0.924	0.804	0.884
CTDT	87.69	0.877	0.218	0.905	0.875	0.676	0.829
DC	93.08	0.931	0.162	0.931	0.929	0.818	0.884
Geary	92.56	0.926	0.157	0.925	0.924	0.804	0.884
Moran	91.54	0.915	0.161	0.914	0.914	0.777	0.877
MoreauBruto	92.05	0.921	0.141	0.92	0.92	0.793	0.89
TC	88.72	0.887	0.308	0.899	0.877	0.703	0.79

Table 1: Classification results using different features for SMO

Fig 2 shows the graphical representation of comparison of classification results using different features for SMO method based on accuracy for the training data. It shows that the highest accuracy of 93.08% is achieved by DC followed by Ctriad which has achieved an accuracy of 92.56%. The lowest accuracy of 87.69% is given by CTDT feature.

Fig 3 shows the classification results using different features for SMO based on precision, F-Measure and MCC for the training data. Thebest precision, F-Measure and MCC valuesof 0.931, 0.929 and 0.818 respectively are provided by DC feature followed by CTriadfeature which gives 0.925, 0.924 and 0.804 values. The lowest values of precision, F-Measure and MCC are provided by CTDT features.

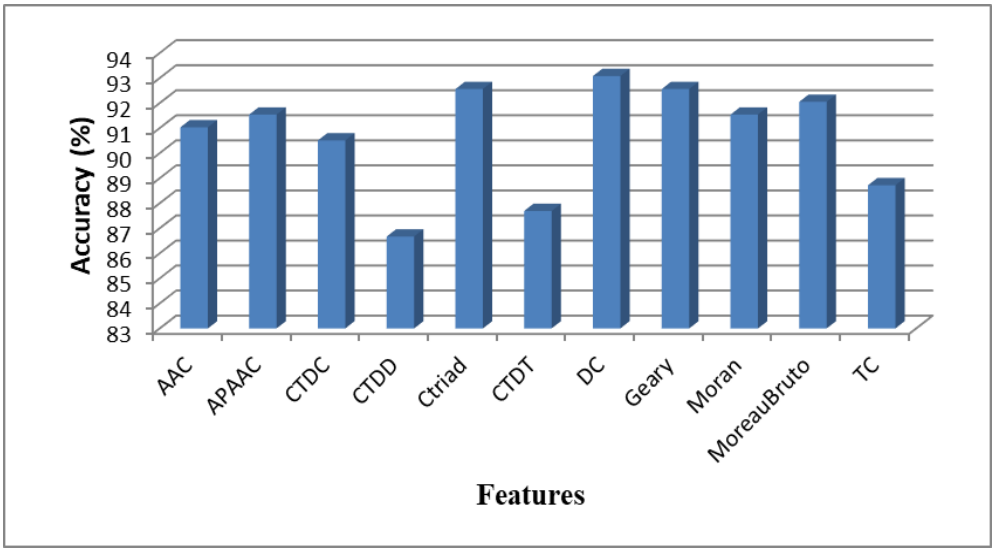

Fig2: Comparison of classification results using different feature for SMO based on accuracy (for training data)

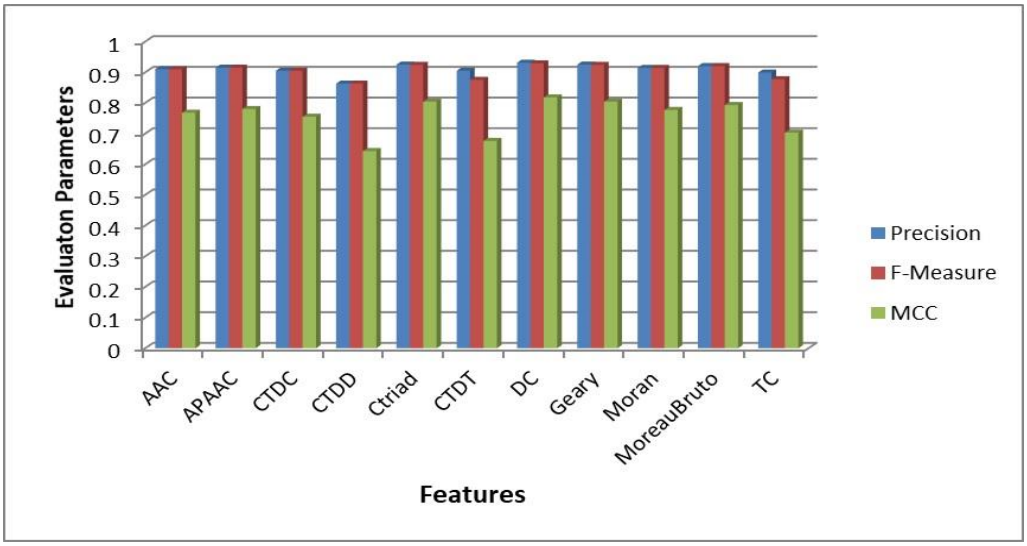

Fig3: Comparison of classification results using different features for SMO based on precision, F-Measure and MCC

Table2 shows the classification results for LIBSVM on training dataset. The highest accuracy is achieved by the optimization of the classifier with the RBF kernel for TC feature which is described as 95.38% accuracy for γ = 2 and C = 150.

Feature	Accuracy (%)	TPR	FPR	Precision	F-Measure	MCC	ROC Area
AAC	90.26	0.903	0.215	0.902	0.899	0.74	0.844
APAAC	88.21	0.882	0.316	0.893	0.871	0.687	0.783
CTDC	91.03	0.91	0.15	0.91	0.91	0.767	0.88

CTDD	84.62	0.846	0.261	0.844	0.845	0.597	0.793
Ctriad	92.82	0.928	0.125	0.928	0.928	0.813	0.901
CTDT	87.95	0.879	0.186	0.88	0.88	0.691	0.847
DC	93.33	0.933	0.117	0.933	0.933	0.827	0.908
Geary	91.28	0.913	0.218	0.916	0.908	0.769	0.847
Moran	91.03	0.91	0.182	0.909	0.908	0.762	0.864
MoreauBruto	93.08	0.931	0.118	0.93	0.953	0.82	0.906
TC	95.38	0.954	0.116	0.955	0.953	0.88	0.919

Table2: Classification results using different features for LIBSVM

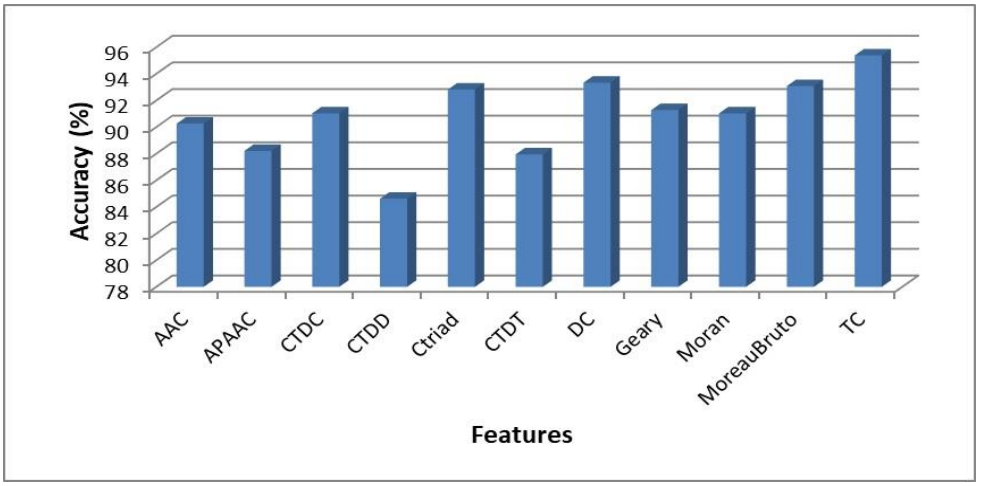

Fig4: Comparison of classification results using different features for LIBSVM based on accuracy

Fig 4 shows the graphical representation of comparison of classification results using different features for LIBSVM method based on accuracy for the training data. It shows that the highest accuracy of 95.38% is achieved by TC followed by DC which has achieved an accuracy of 93.33%. The lowest accuracy of 84.69% is given by CTDT feature. Fig 5 shows the classification results using different features for LIBSVM based on precision, F-Measure and MCC for training data. The best precision, F-measure and MCC values of 0.955, 0.953 and 0.88 respectively are provided by TC feature followed by DC feature which gives 0.933, 0.933 and 0.827 values. The lowest value of precision, F-measure and MCC are provided by CTDT.

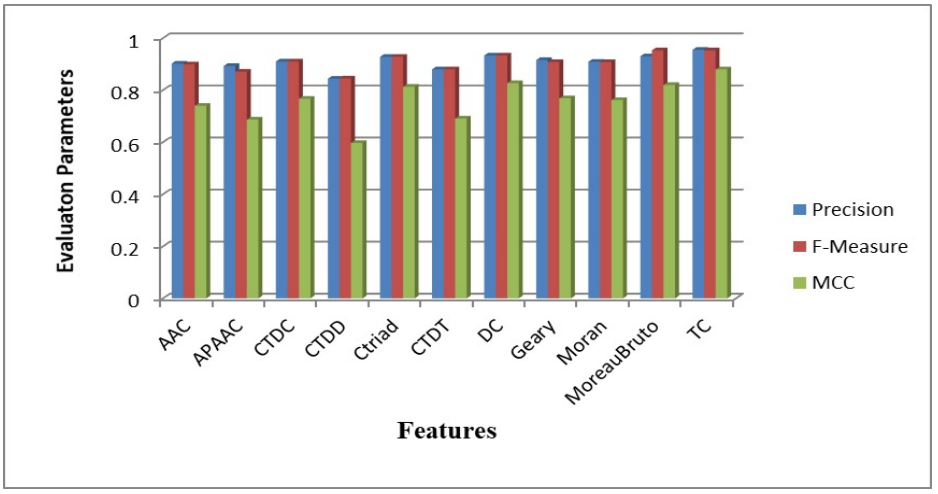

Fig 5: Comparison of classification results using different features for LIBSVM based on precision, F-Measure and MCC (for training data).

Table3 shows the Classification results for RF on training dataset. The highest accuracy is achieved by the optimization of the classifier for AAC feature.

Feature	Accuracy (%)	TPR	FPR	Precision	F-Measure	MCC	ROC Area
AAC	91.79	0.918	0.154	0.917	0.917	0.785	0.97
APAAC	90.77	0.908	0.207	0.907	0.904	0.754	0.942
CTDC	86.92	0.869	0.259	0.865	0.865	0.649	0.912
CTDD	88.72	0.887	0.221	0.885	0.884	0.7	0.929
Ctriad	86.67	0.867	0.284	0.863	0.861	0.639	0.904
CTDT	86.67	0.867	0.284	0.863	0.861	0.639	0.904
DC	90.77	0.993	0.245	0.914	0.902	0.757	0.975
Geary	85.64	0.856	0.4	0.88	0.836	0.618	0.93
Moran	85.38	0.854	0.401	0.874	0.834	0.608	0.941
MoreauBruto	87.44	0.874	0.35	0.893	0.86	0.669	0.945
TC	85.13	0.851	0.408	0.872	0.83	0.6	0.945

Table3: Classification results using different features for RF.

Fig 6 shows the graphical representation of comparison of classification results using different features for RF method based on accuracy for the training data. It shows that the highest accuracy of 91.79% is achieved by AAC followed by APAAC which has achieved an accuracy of 90.77%. The lowest accuracy of 85.13% is given by TC feature. Fig 7 shows the classification results using different features for RF based on precision, F-Measure and MCC for training data. The best precision, F-measure and MCC value of 0.917, 0.917 and 0.785 respectively are provided by AAC feature followed by APAAC feature which gives 0.907, 0.904 and 0.754 values. The lowest value of precision, F-measure and MCC are provided by TC.

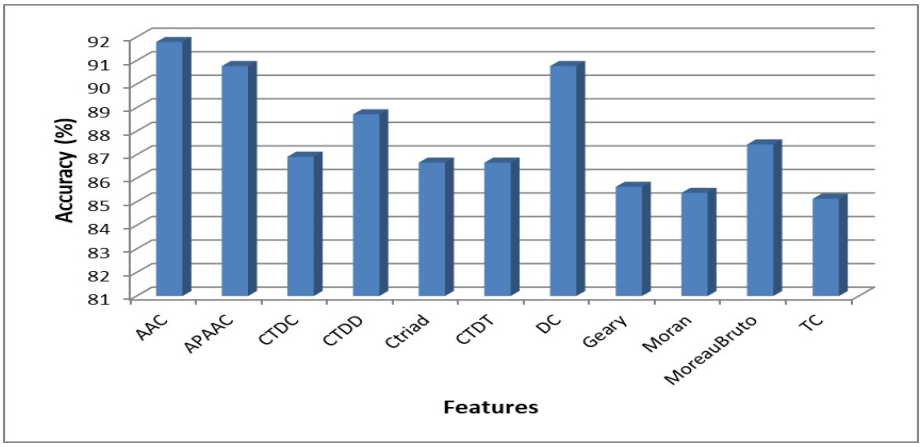

Fig 6: Comparison of classification results using different features for RF based on accuracy

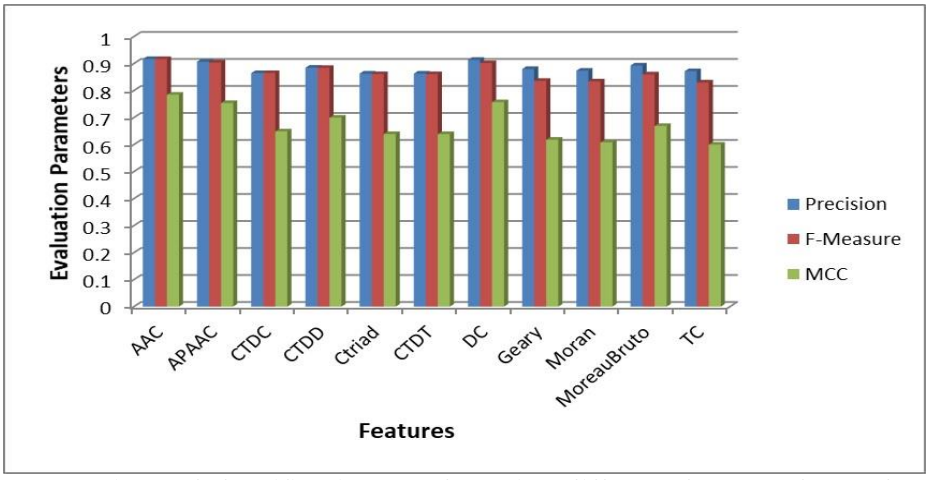

Fig 7: Comparison of classification results using different features for RF based on precision, F-Measure and MCC

Classification Results for the dataset for testing data

This sub-section presents the classification results on testing data using the best feature obtained from the classification results on the training data. Table 7 shows the classification results of all the three classifiers i.e., SMO, LIBSVM and RF for test data. The highest accuracy of 90.91% is achieved by the RF for AAC feature. SMO with RBF kernel provides the second highest accuracy of 88.89% for $\gamma = 0.09$ and $C = 250$ for the feature DC. The lowest accuracy is achieved by LIBSVM for TC feature.

Table 7: Classification results (for testing data) using the best feature.

Classifier	Best feature	Accuracy (%)	TPR	FPR	Precision	F-Measure	MCC	ROC Area
SMO	DC	88.89	0.889	0.262	0.889	0.883	0.7	0.813
LIBSVM	TC	83.84	0.838	0.429	0.851	0.816	0.552	0.705

RF	AAC	90.91	0.909	0.156	0.908	0.909	0.763	0.94

ROC curves illustrate the trade-off between TPR (sensitivity) and FPR (1-specificity) over their whole range of feasible values. It is considered as the most robust approach for classifier evaluation(Viscaino & Cheein, 2019). The Area under Curve (AUC) is used as a consistent index of classifier performance(Zhang, 2019). This validates the threshold-independent performance of the classifiers (Hueso et al., 2018). Fig 14 shows the ROC curve for the best classifier i.e., RF classifier for AAC feature for the testing data .Its AUC is 0.94 which auxiliary reinforces the discriminative efficiency of the model.

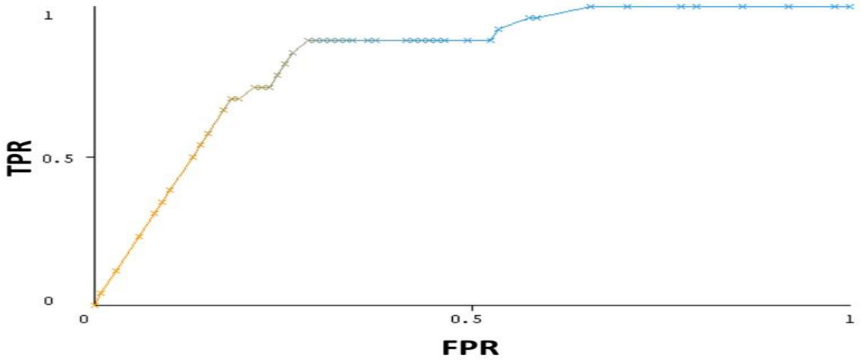

Fig 14: ROC plot for AAC feature obtained from RF Classifier

The results show that the performance quality of our method concerning accuracy, precision, and four other metrics, as well as the results obtained, are acceptable and encouraging. The method offers a promising approach for the prediction of NLR protein. The goal of this study is to further augment the knowledge about NLR protein which will lead to several other strategies for controlling the microbiota within the intestine and in between digestive issues. The identification method developed in this work can expedite the discovery of NLR proteins and needs to be judiciously used. The model generated can be used to identify NLR protein across various organisms. We hope this would be a helpful method for NLR prediction to the end-user biologist and the research community as a whole.

Conclusion

This paper presented a new method named as NLRPred for the identification of NLR from other proteins with high selectivity. NLRPred identification results prove that the method may be used as an automatic tool for the identification of NLR. The study has carried out a comparison of three classification methods i.e., SMO, LIBSVM and RF for the prediction of NLR proteins. The dataset used for this work is created after extracting the features using ProtR package. The best classification feature is identified for the training data. After analyzing the training data results for both the packages, the best feature obtained is employed on the testing data. Overall, AAC based model from RF classifier was able to achieve an accuracy of 90.91%. Thus, RF proved to be the best algorithm for

prediction of NLR protein. In future, the new sequences and their relationships will be explored to include the new set of attributes to improve the classification performance.

List of Abbreviations

SVM, Support Vector Machine; PSSM, Position-specific scoring matrix; AAC, Amino acid compositions; RBF, Radial basis function; ROC, Receiver Operator Characteristic; MCC,

Conflict of interest

The authors declare no conflict of interest, financial or otherwise.

Acknowledgements

We thank Jayashree Ramana who conceptualized the project and provided continual guidance and discussion. We would also like to thank JUIT for providing us with the infrastructure and all the facilities required to carry out this work.

References

1. Adachi, H., Contreras, M. P., Harant, A., Wu, C.-H., Derevnina, L., Sakai, T., . . . Kamoun, S. (2019). An N-terminal motif in NLR immune receptors is functionally conserved across distantly related plant species. *eLife, 8*, e49956. doi:10.7554/eLife.49956
2. Agius, R., Brieghel, C., & Andersen, M. A. (2020). Machine learning can identify newly diagnosed patients with CLL at high risk of infection. *11*(1), 363. doi:10.1038/s41467-019-14225-8
3. Amouri, A., Alaparthy, V. T., & Morgera, S. D. (2020). A Machine Learning Based Intrusion Detection System for Mobile Internet of Things. *Sensors (Basel), 20*(2). doi:10.3390/s20020461
4. Baggs, E., Dagdas, G., & Krasileva, K. V. (2017). NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. *Current Opinion in Plant Biology, 38*, 59-67. doi:https://doi.org/10.1016/j.pbi.2017.04.012
5. Biswas, A., & Kobayashi, K. S. (2013). Regulation of intestinal microbiota by the NLR protein family. *Int Immunol, 25*(4), 207-214. doi:10.1093/intimm/dxs116
6. Chen, G. Y. (2014). Role of Nlrp6 and Nlrp12 in the maintenance of intestinal homeostasis. *Eur J Immunol, 44*(2), 321-327. doi:10.1002/eji.201344135
7. Fletcher, R. R., Olubeko, O., Sonthalia, H., Kateera, F., Nkurunziza, T., Ashby, J. L., . . . Hedt-Gauthier, B. (2019). Application of Machine Learning to Prediction of Surgical Site Infection. *Annu Int Conf IEEE Eng Med Biol Soc, 2019*, 2234-2237. doi:10.1109/embc.2019.8857942
8. Hartmann, S., & Baumert, M. (2019). Improved A-phase Detection of Cyclic Alternating Pattern Using Deep Learning. *Annu Int Conf IEEE Eng Med Biol Soc, 2019*, 1842-1845. doi:10.1109/embc.2019.8857006
9. Higashi, K., Sun, G., & Ishibashi, K. (2019). Precise Heart Rate Measurement Using Non-contact Doppler Radar Assisted by Machine-Learning-Based Sleep

Posture Estimation. *Annu Int Conf IEEE Eng Med Biol Soc, 2019*, 788-791. doi:10.1109/embc.2019.8857830

10. Hirota, S. A., Ng, J., Lueng, A., Khajah, M., Parhar, K., Li, Y., . . . Beck, P. L. (2011). NLRP3 inflammasome plays a key role in the regulation of intestinal homeostasis. *Inflammatory Bowel Diseases, 17*(6), 1359-1372. doi:10.1002/ibd.21478
11. Hueso, M., Vellido, A., Montero, N., Barbieri, C., Ramos, R., Angoso, M., . . . Jonsson, A. (2018). Artificial Intelligence for the Artificial Kidney: Pointers to the Future of a Personalized Hemodialysis Therapy. *Kidney Dis (Basel), 4*(1), 1-9. doi:10.1159/000486394
12. Jagga, Z., & Gupta, D. (2014). Supervised learning classification models for prediction of plant virus encoded RNA silencing suppressors. *PLoS ONE, 9*(5), e97446. doi:10.1371/journal.pone.0097446
13. Jagga, Z., & Gupta, D. (2015). Machine learning for biomarker identification in cancer research - developments toward its clinical application. *Per Med, 12*(4), 371-387. doi:10.2217/pme.15.5
14. JM, O. T., & Boylan, G. B. (2019). Machine learning without a feature set for detecting bursts in the EEG of preterm infants. *Annu Int Conf IEEE Eng Med Biol Soc, 2019*, 5799-5802. doi:10.1109/embc.2019.8856533
15. Kalita, M. K., Nandal, U. K., Pattnaik, A., Sivalingam, A., Ramasamy, G., Kumar, M., . . . Gupta, D. (2008). CyclinPred: a SVM-based method for predicting cyclin protein sequences. *PLoS ONE, 3*(7), e2605. doi:10.1371/journal.pone.0002605
16. Kigka, V. I., Sakellarios, A. I., Tsompou, P., Kyriakidis, S., Siogkas, P., Andrikos, I., . . . Fotiadis, D. I. (2019). Site specific prediction of atherosclerotic plaque progression using computational biomechanics and machine learning. *Annu Int Conf IEEE Eng Med Biol Soc, 2019*, 6998-7001. doi:10.1109/embc.2019.8856881
17. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J, 13*, 8-17. doi:10.1016/j.csbj.2014.11.005
18. Kumar, M., Gromiha, M. M., & Raghava, G. P. (2008). Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins, 71*(1), 189-194. doi:10.1002/prot.21677
19. Levy, A., Stedman, A., Deutsch, E., Donnadieu, F., & Virgin, H. W. (2020). Innate immune receptor NOD2 mediates LGR5(+) intestinal stem cell protection against ROS cytotoxicity via mitophagy stimulation. *117*(4), 1994-2003. doi:10.1073/pnas.1902788117
20. Liao, L., & Schneider, K. M. (2019). Intestinal dysbiosis augments liver disease progression via NLRP3 in a murine model of primary sclerosing cholangitis. *68*(8), 1477-1492. doi:10.1136/gutjnl-2018-316670
21. Molteni, E., Colombo, K., Beretta, E., Galbiati, S., Santos Canas, L. D., Modat, M., & Strazzer, S. (2019). Comparison of Multi-class Machine Learning Methods for the Identification of Factors Most Predictive of Prognosis in Neurobehavioral assessment of Pediatric Severe Disorder of Consciousness through LOCFAS scale. *Annu Int Conf IEEE Eng Med Biol Soc, 2019*, 269-272. doi:10.1109/embc.2019.8856880

22. Nadia, & Jayashree, R. (2020). The Human OncoBiome Database: A Database of Cancer Microbiome Datasets. *Current Bioinformatics, 15*(5), 472-477. doi:http://dx.doi.org/10.2174/1574893614666190902152727
23. Nudel, J., Bishara, A. M., de Geus, S. W. L., Patil, P., Srinivasan, J., Hess, D. T., & Woodson, J. (2021). Development and validation of machine learning models to predict gastrointestinal leak and venous thromboembolism after weight loss surgery: an analysis of the MBSAQIP database. *Surg Endosc, 35*(1), 182-191. doi:10.1007/s00464-020-07378-x
24. Ramana, J., & Gupta, D. (2009). LipocalinPred: a SVM-based method for prediction of lipocalins. *BMC Bioinformatics, 10*, 445. doi:10.1186/1471-2105-10-445
25. Ramana, J., & Gupta, D. (2010a). FaaPred: a SVM-based prediction method for fungal adhesins and adhesin-like proteins. *PLoS ONE, 5*(3), e9695. doi:10.1371/journal.pone.0009695
26. Ramana, J., & Gupta, D. (2010b). Machine learning methods for prediction of CDK-inhibitors. *PLoS ONE, 5*(10), e13357. doi:10.1371/journal.pone.0013357
27. Seo, S. U., Kamada, N., Muñoz-Planillo, R., Kim, Y. G., Kim, D., Koizumi, Y., . . . Núñez, G. (2015). Distinct Commensals Induce Interleukin-1β via NLRP3 Inflammasome in Inflammatory Monocytes to Promote Intestinal Inflammation in Response to Injury. *Immunity, 42*(4), 744-755. doi:10.1016/j.immuni.2015.03.004
28. Tamanna, & Ramana, J. (2015). MATEPRED-A-SVM-Based Prediction Method for Multidrug And Toxin Extrusion (MATE) Proteins. *Comput Biol Chem, 58*, 199-204. doi:10.1016/j.compbiolchem.2015.07.011
29. Tiwari, P., Colborn, K. L., Smith, D. E., Xing, F., Ghosh, D., & Rosenberg, M. A. (2020). Assessment of a Machine Learning Model Applied to Harmonized Electronic Health Record Data for the Prediction of Incident Atrial Fibrillation. *JAMA Netw Open, 3*(1), e1919396. doi:10.1001/jamanetworkopen.2019.19396
30. Viscaino, M., & Cheein, F. A. (2019). Machine learning for computer-aided polyp detection using wavelets and content-based image. *Annu Int Conf IEEE Eng Med Biol Soc, 2019*, 961-965. doi:10.1109/embc.2019.8857831
31. Xu, L., Liang, G., Liao, C., Chen, G. D., & Chang, C. C. (2018). An Efficient Classifier for Alzheimer's Disease Genes Identification. *Molecules, 23*(12). doi:10.3390/molecules23123140
32. Zahid, A., Li, B., Kombe, A. J. K., Jin, T., & Tao, J. (2019). Pharmacological Inhibitors of the NLRP3 Inflammasome. *Front Immunol, 10*, 2538. doi:10.3389/fimmu.2019.02538
33. Zhang, L. (2019). EEG Signals Classification Using Machine Learning for The Identification and Diagnosis of Schizophrenia. *Annu Int Conf IEEE Eng Med Biol Soc, 2019*, 4521-4524. doi:10.1109/embc.2019.8857946