Journal of Mountain Area Research

Vol. 8, 2023

https://doi.org/10.53874/jmar.v8i0.166

Full length article

IDENTIFICATION MOLECULAR FUNCTIONS OF DYNEIN MOTOR PROTEINS USING EXTREME GRADIENT BOOSTING ALGORITHM WITH MACHINE LEARNING

Ali Ghulam[1], Rahu Sikander[2], Dhani Bux Talpur[3], Erum Saba[1], Mir Sajjad Hussain Talpur[1], Zulfikar Ahmed Maher[1], Saima Tunio[1]*

1. Information Technology Centre, Sindh Agriculture University, Sindh, Pakistan
2. School of Computer Science and Technology, Xidian University, Xi'an 710071, China
3. Department of Computer Science, University of Gwaddar, Gwaddar, Balochistan

ABSTRACT

The majority of cytoplasmic proteins and vesicles move actively primarily to dynein motor proteins, which are the cause of muscle contraction. Moreover, identifying how dynein are used in cells will rely on structural knowledge. Cytoskeletal motor proteins have different molecular roles and structures, and they belong to three superfamilies of dynamin, actin and myosin. Loss of function of specific molecular motor proteins can be attributed to a number of human diseases, such as Charcot-Charcot-Dystrophy and kidney disease. It is crucial to create a precise model to identify dynein motor proteins in order to aid scientists in understanding their molecular role and designing therapeutic targets based on their influence on human disease. Therefore, we develop an accurate and efficient computational methodology is highly desired, especially when using cutting-edge machine learning methods. In this article, we proposed a machine learning-based superfamily of cytoskeletal motor protein locations prediction method called extreme gradient boosting (XGBoost). We get the initial feature set All by extraction the protein features from the sequence and evolutionary data of the amino acid residues named BLOUSM62. Through our successful eXtreme gradient boosting (XGBoost), accuracy score 0.8676%, Precision score 0.8768%, Sensitivity score 0.760%, Specificity score 0.9752% and MCC score 0.7536%. Our method has demonstrated substantial improvements in the performance of many of the evaluation parameters compared to other state-of-the-art methods. This study offers an effective model for the classification of dynein proteins and lays a foundation for further research to improve the efficiency of protein functional classification.

KEYWORDS: Dynein motor proteins, Machine learning, BLOUSM62, Computational Methods

*Corresponding author: (Email: garahu@sau.edu.pk)

1. INTRODUCTION

Muscle contraction is driven by motor proteins, and proteins and vesicles play an important role in cytoplasmic transport. Chemical energy from the hydrolysis of adenosine triphosphate (ATP) converts these proteins into mechanical tasks that move along actin filaments or microtubules. Adenosine triphosphate (ATP) hydrolysis' chemical energy can be converted by these proteins into mechanical work that flows along actin filaments or microtubules. To provide the mechanical forces needed to power biological movement, various methods have developed. Mechanochemical enzymes, or "motor proteins," are a highly effective and common method of producing biological force [1, 2]. Myosin motors operating on actin filaments cause muscle cell contraction, vesicle movement, cytoplasmic streaming, and other morphological changes. Members of the dynein and kinesin microtubule-based motor superfamilies move vesicles and organelles inside of cells, cause the beating of flagella and cilia, and segregate replicated chromosomes to offspring cells in the mitotic and meiotic spindles. [3]. Transport of different freight products, including membrane organelles, protein complexes, and mRNAs directly involved in the dynein superfamily of proteins [4]. The dynein and kinesin superfamilies of microtubule motors are responsible for vesicle and organelle movement within cells, which causes flagella and cilia to beat. They also work within mitotic and meiotic spindles to identify duplicated chromosomes. [4]. For example, dynamin and cytoplasmic dynamin can be detected in spinal cord spheres associated with motor neuron disease [5]. Neurodegenerative diseases are also associated with mutations in dynein proteins and targeted disruption of function [6, 7]. T cell-mediated disease [8] is myosin-induced acute myocarditis. Myosin variant Ixb increases the risk of celiac disease, leading to primary intestinal barrier defects [9].

Many bioinformatics researchers are interested in the important role that motor proteins play in human function. For example, Zhu, C et al. [10] have established an overview of the factor superfamily of dynein motor proteins, in which they focus on the structure and function of motor dynein proteins. The bioinformatics system for classifying heavy dynamin chains was subsequently introduced by Yagi [11]. Khataee and Liew [12] projected a computational model for studying forward and reverse dynamics on the basis of the four-state discrete random motion model. The molecular function of motor dynein protein consists of a double-stranded dimer. The polypeptides associated with each feature (shown in pink) are different and a complex set of intermediary chains, light intermediate chains, and light chains in dynamin). (Readers are referred to the online version of this article to explain the color references in this legend.) A process simulation model for dynein [13]. Stedman et al. [14] proved the association between dynein gene mutation and human anatomical lineage changes by using bioinformatics method. Some previous studies in bioinformatics have also been used to detect dynein activity and dynein phosphatase diversity [15, 16].

We proposed computational protein sequence analysis, we create molecular level biological problems at the level of single molecules, cells, and tissues in this work. We used machine learning algorithm for the prediction of accuracy, sensitivity, specity and then MCC, live-cell imaging in neurons, and computational modelling to analyses the regulation of the beginning of dynein-dynactin transport. In some bioinformatics researchers, WEKA is an automatic learning method for data mining technology [17]. Secondly, various problems related to protein functional classification were obtained by using RBF network [18, 19], good results have been obtained. In addition, LibSVM [20] the emergence of deep learning means that the field of bioinformatics must become more efficient. Machine learning is an advanced computer education method, which uses artificial intelligence to learn representative data with multiple neural network layers [21]. There are many benefits to using machine learning, such as getting the latest results, reducing the need for extraction functions, and so on (GPU). In recent years, bioinformatics has shifted from traditional machine learning and protein function learning to in-depth learning strategies. For example, as tried by Alipanahi et al., DNA and RNA proteins learn the sequence properties of the binding in depth. Spencer et al. [22] proposed an ab initio deep learning network for protein secondary structure prediction. To solve this problem, [23] create Stack-ACPred, a brand-new predictor for accurately identifying ACPs. To construct the stacking-base ensemble model for targeting efficient ACPs, the best qualities should be chosen. [24] suggested bioinformatics technique thus outperformed all current state-of-the-art sequence based CPP methods. In the current study, [25] developed a novel feature extraction technique that takes into account the influence of residues in the vicinity of the mutation site. To assess the effectiveness of the suggested feature extracting method, rigorous cross-validation and independent tests were run on benchmark datasets. The molecular processes that enable these numerous and varied tasks are the subject of this research. Therefore, it is crucial to increase the immunoglobulin classification's accuracy by using efficient illness research techniques. Based on the BLOSUM vector score matrix, we extract IgG characteristics using the reduced feature dimension features that were chosen. Extreme Gradient Boosting is an ensemble learning technique we've created (XGBoost) [26]. Cancer peptide therapy is interesting since it offers so many alluring advantages. Anticancer Peptides (ACPs), which are essential for the development of innovative cancer therapeutics, have attracted a lot of attention from researchers in recent years. Experimental methods are expensive, time-consuming, and frequently produce unreliable predictions when used to forecast ACPs [27]. In order to increase the linkages between disease variation and new molecular correlations between genetic mutations, we carried out a pathway-based investigation [28]. On the basis of shared gene interactions among illness-pathways, we created a biological network, and then we used network analysis to try and understand how a disease develops.

In this article, we suggested dynein motor proteins with XGBoost algorithm to predict effectors using attributes deduced from sequencing. As a result, this study suggests using an XGBoost algorithm built from a position-specific scoring matrix. We judiciously summarized the elements addressed in other research and added a few new features to get around the shortcomings of the current methodologies. As shown in a series of recent publications [29], the guidelines under the five-step rule should be followed to create a truly useful sequence-based statistical prediction for a biological or biomedical approach. We would like to explain the five steps here: how I should construct the biological sequence of samples to generate, select, or predict a valid data set that can realistically represent their interactions with the predicted target. How the biological sequence of samples should be formed or built to run a powerful algorithm (or engine), we will explain how these measures are accomplished one by one below.

Table 1. Statistics of all retrieved dynein proteins.

	Original data	Similarity <30%	Cross-validation
Dynein protein	4153	721	306
non-Dynein protein	4153	721	252
Total Protein	8306	1442	558

2. MATERIALS AND METHODS

We proposed a novel computational method to analyze the molecular activity of high-performance dynein protein using eXtreme gradient boosting (XGBoost) firstly Dataset Collection, Feature Extraction, and Model Interpretation make up the two stages of the dynein protein-XGB procedure, which is depicted in Figure 1. The following sections provide a full description of each stage.

2.1 Data collection

In our study, we chose secreted effectors and non-effectors to create the benchmark dataset and build the machine learning-based model for dynein protein prediction. Our dataset, which included 250 non-dynein proteins and 520 dynein proteins, was directly retrieved from the recently published work (Wang, et al., 2019). The CD-HIT programmed ran the protein dataset through a filter of >30% sequence identity to cut down on sequence redundancy [30]. A total of 400 dynein proteins and 200 non-dynein proteins made up the final training dataset, and 120 dynein proteins and 50 non-dynein proteins made up the independent test dataset.

$$I = I^{+} \cup I^{-} \qquad (1)$$

The subjects of this study were positive dynein protein 400 examples of good behavior made up the participants of this study's I+ dataset, while 230 examples of negative behavior made up the target population of the I dataset. In conclusion, Table 1 shows that the baseline dataset consists of 228 protein sequences. We employed two different data sets for independent validation in order to further highlight the accuracy of the technique discussed in this study.

2.2 Feature extraction technique

In this work, we identified the coevolution relationship using 3080 dynein motor protein characteristics MSAs. At least one family will have a structure that has been determined empirically, and each family will include sequences of 100, 400, and 700 lengths. The "no gap" location is the only one the MSA will calculate. A "no-gap" position, then, is one that occupies less than 10% of the clearance. We employed mutual information to extract the association between evolutionary coevolution. We extracted features sequence data of the protein is represented by the BLOSUM62 matrix profiles. The matrix with m' L elements, where L is the length and m = 20 is the number of amino acids, serves as a representation for each residue in the training dataset. The 20 frequent amino acids are represented by one row in the normalized BLOSUM62 matrix. Equal-length peptides can be encoded with the BLOSUM62 descriptor. The BLOSUM62 matrix profiles, a regularly used tool for determining the alignment of two different protein sequences, was used in this work. By analyzing observed polypeptide alignments on a vast scale, the value of the BLOSUM62 Matrix is calculated. In MSA, mutual information between two locations is described as information like [31] between two sites.

$$L_{ij} = \Sigma^{a}_{A_i,B_j} f\left(A_i,B_j\right) log \frac{f\left(A_i,B_j\right)}{f\left(A_i\right) f\left(B_j\right)} \quad (2)$$

where, more particularly, we used 21 different types of amino acids, calculating the gap as the 21st amino acid, if q is equal to 21. Although there are enough variables in the MSA sequence data set, we use frequency to roughly approximation the likelihood. The frequency of a single type A amino acid seen at position I is indicated by f (Ai). The frequency f of the two types of amino acids occurring at locations I and J combined f (Ai, Bj). Sequence redundancy must be taken into account when calculating single- and dual-frequency weights. We compute all the mutual information in reference [32].

When assessing coevolution correlations, a false positive can also happen, as was the case in the current study. We used the structural data that was present in the training data set. Remainders are only paired together if they are architecturally close to one another. The "adjacent residues" are defined as those residues with tip atoms that are separated by 4.5 angstroms or less. A group of residues are more likely to interact with one another the more dissimilar their sequences are and the closer they are to one another.

2.3 Derive a new matrix from contextual substitutions

An element in the amino acid substitution matrix that corresponds to the substitution percent was used to categories the target population. How frequently one amino acid gets swapped out for another is indicated by the substitution fractions. Coevolution calculations show that there are universal relationships between protein families. Instead of happening in isolation, most amino acid changes take place in concert with their structural context. In addition, alternate fractions that are identical to the original BLOSUM vector matrix are produced using the logarithmic odds ratio [33].

$$s\left(A,B\right) = \lambda log \frac{f\left(A,B\right)}{f\left(A\right) f\left(B\right)} \quad (3)$$

In this section A and B must be substituted for each other in order to determine the joint frequency F (A, B) using the expected frequency for a single amino acid type and the actual joint frequency. Is a scalar factor whose objective is to succeed.

2.4 Model Proposed

The goal of the current study was to accurately predict if a protein sequence is that of a dynein protein; this categorization issue is known as a dichotomy problem [34, 35]. Three classifiers were used in this paper's analyses in order to choose those that could predict dynein proteins more precisely than the others. SVM, AdaBoost, KNN, and Random Forest (RF) were the three classifiers employed. This observation might be

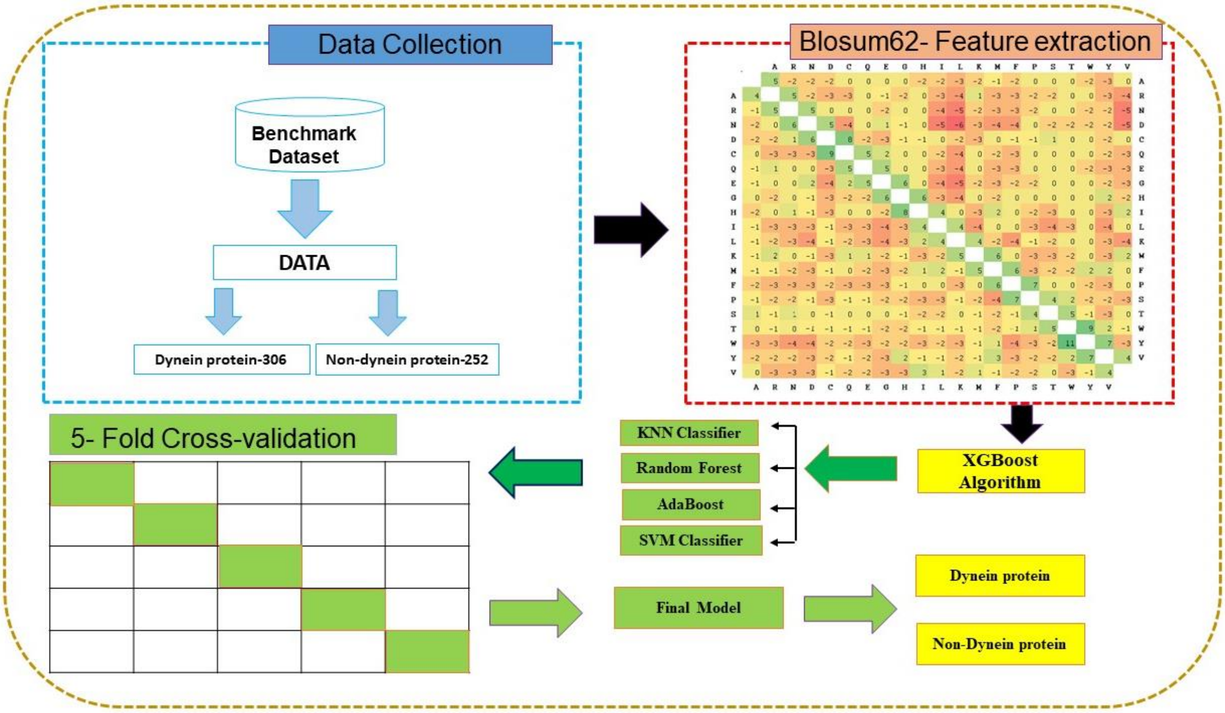

Figure 1. Proposed framework model

interpreted as showing that a nuclear's solvent accessibility is determined by the amount of atomic surface that is exposed to solvents. The degree of atomic accessibility in a residue determines its residue availability [36]. Similarly, residues are regarded as exposed residues if their solvent accessibility is higher than the threshold. The following diagram illustrates several stages of the dataset processing, function formulation, RF model generation, and novel protein sequence prediction processes.

2.5 Model Evaluation performance

In this work, to evaluate how well the classifiers employed in this work performed, the following metrics were used. Accuracy, sensitivity, F-measure, Matthew's correlation coefficient (MCC), and area under receiver operating characteristic curve (AUC) [37, 38] were used to evaluate the performance of the classifiers utilized in this investigation. As the prediction threshold was changed, a trade-off between sensitivity and specificity was seen. AUC offers a single measure of total threshold-independent accuracy, making it a useful tool for comparing the overall prediction performance of various approaches. Perfect prediction accuracy is shown by an AUC and MCC of 1. As follows is a definition of these measures:

3. RESULTS AND DISCUSSION

In this part, we used the 5-fold cross validation method on the training data set to assess the prediction strength of the various feature categories individually and in combination. The training data set was partitioned into five subsets at random for 5-fold cross validation. Four subsets were utilized to train XGBoost, and the final one was used to assess the model's effectiveness. Five times each of the processes were repeated. ACC and dynein proteins, two performance indicators from the training set, were averaged, and the findings are displayed in Table 1. On the training data set, it can be shown that some specific Fasta feature categories have a greater overall prediction power. This finding suggests that, when compared to other types of characteristics, Fasta-based features perform better in the prediction of the dynein protein. We perfumed. The combining various features could provide a more thorough representation of protein sequences [39, 40]. The combined characteristics, as shown in Table 2, produce the ACC of 93.95 percent and the MCC of 0.8346, both of which are greater than other Fasta-based features. In conclusion, the combination of all characteristics regularly outperformed single feature-based models in terms of performance.

$$Sensitivity = \frac{TP}{TP+FN} \tag{4}$$

$$Specificity = \frac{TN}{TN+FP} \tag{5}$$

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{6}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{7}$$

Information about the amounts of h dynein protein and non-dynein protein in the aforementioned formulations is provided (true negative). We calculated false negative and false positive results are also regarded as non-dynein proteins. In addition, utilizing X-axis sensitivity and Y-axis accuracy, we discovered an operator receptor curve (ROC). When assessing the model effectiveness of different choice values, the significance of the AUC is important. Horses and metaphors are both equally useful, accurate, and adaptable. The connections between accuracy, precision, sensitivity, specificity and MCC score have been identified in order to find the correlation of Matthew similarities.

3.1 Impact of BLOSUM62 Extraction Algorithm performance

Building novel protein sequences should take into account the difficulty of obtaining random protein sequences using eXtreme gradient boosting (XGBoost), Blossom62, due to the high synthesis of dynein protein . For several categories, the same data set is utilized. This is the only effective method for reducing the function's mixed space (Blossom62). The value of the ensuing Spaces is shown by the existence of these varied individual and mixed places. Analysis using random forests was done. The ideal model growth parameter is discovered via the eXtreme gradient boosting (XGBoost), Blossom62 strategy. Content analysis is utilized as an indicator for the analysis of parameters impacting the dynein protein and to gauge how well the dynein protein is functioning. The accordance with maintaining consistent precision values between dynein protein structures, and their corresponding dynein protein precision values are displayed in Table 2.

Table 2. eXtreme gradient boosting (XGBoost), identifying optimum parameter for various models

	ACC	Precision	Sensitivity	Specificity
XGBoost	0.8676%	0.8768%	0.760%	0.9752%

The analysis was based on the results of a 5-fold cross-validation of our suggested models using the properties of the BLOSUM vector matrix. Fivefold cross-validation was used to evaluate the performance of our proposed model on both positive and negative data sets. We trained the model based on training sets we employed in our experiment, we analyses the outcomes. We achieved prediction performance was relatively strong, with an overall accuracy of 0.8676% percent, according to the computational study. Figure 2 displays the ROC (AUC) curve score that we were given. As seen by the predicted score and the ROC-AUC of 0.91% percent, our model's estimation is quite accurate. The dynein proteins exhibit strong ROC curve performance (auc=0.91% percent), and the ROC curve also works well, using a BLOSUM -extracting feature model.

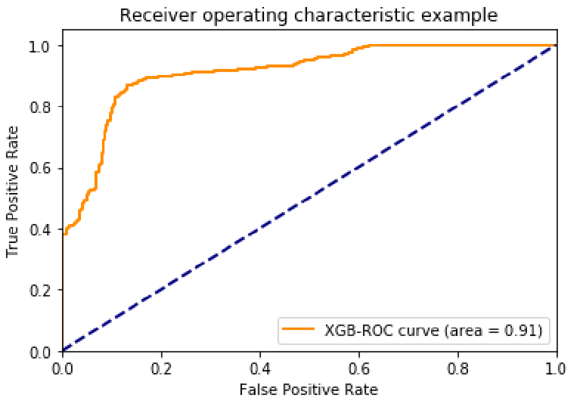

Figure. 2. Proposed ROC-AUC score predictive models

3.2 Proposed methods comparison performance with other ML classifiers

The objective was to compare the performance of the algorithms Adaboost, RF, SVM, and K-NN, four widely used techniques in the field, in order to demonstrate eXtreme gradient boosting (XGBoost) superiority in dynein protein sequence prediction. The training technique was created with Python 3.8, finished with 5-fold cross-validation, and included relatively extensive parameter tuning. We investigation support as shown Figure 3 validity intuitively. Processed by dynein proteins the precision of AUC values is 0.92 percent when using the ROC-AUC curve, which employs ROC AUC curves comparable to those used in other methodologies. Three classification systems, including the KNN classifier (KNN), the SVC classifier, and the Random Forest Classifier (RFC), are used to supplement the results for our proposed eXtreme gradient boosting (XGBoost) model [41]. XGBoost, a traditional ensemble learning technique, was used to evaluate KNN and SVM

3.3 Comparison performance with other 4 ML Classifiers hybrid features

Our findings can be compared to results of earlier studies that, we compared performance with the reference models (AdaBoost, RF, SVM, and KNN) based on accuracy, precision, sensitivity, specificity and MCC. We obtained the score based on parameters were properly measured, as shown in Table 3. Excellent results were obtained in this combined comparison with 3080 features length using eXtreme gradient boosting (XGBoost) classification. Following, they employed a data set with features from the BLOSUM-62 vector score in an

.

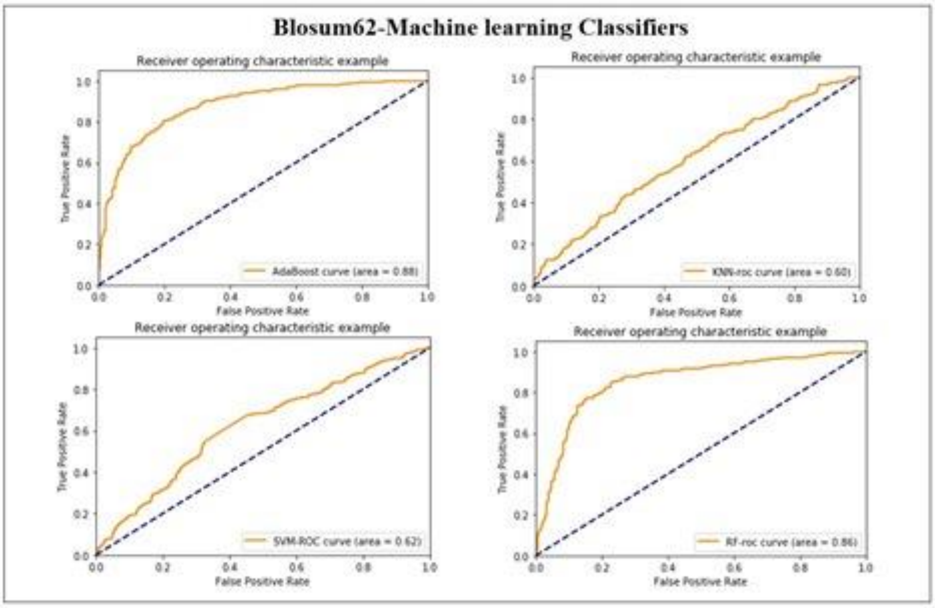

Figure 3. Proposed ROC-AUC score predictive models

eXtreme gradient boosting (XGBoost) model. With BLOSUM retrieved features, we achieved an accuracy score of 0.8676% percent. It is evident that standard machine learning algorithms are outperformed by eXtreme gradient boosting (XGBoost) techniques in terms of prediction accuracy and AUC values.

Table 3. Performance with other 4 ML Classifiers hybrid features

ML-Classifiers	ACC	Precision	Sensitivity	Specificity
AdaBoost	0.8030%	0.7810%	0.775%	0.8311%
KNN	0.5666%	0.5601%	0.6066%	0.5266%
RF	0.8004%	0.7855%	0.775%	0.8259%
SVM	0.6079%	0.6460%	0.4976%	0.7183%
XGBoost	0.8676%	0.8768%	0.76%	0.9752%

3.4 ROC (AUC) comparison of combined 4 ML Classifiers

The data provide convincing evidence the previous article described how this experiment compared the efficacy of four classifiers: AdaBoost, RF, SVM, and KNN. The parameters for the four methods were taken from the classifiers' default settings. The linear SVM kernel by default had consequence coefficient values of C = 1.0. In comparison to the findings of earlier studies, we evaluated the accuracy, precision, sensitivity, specificity and MCC of the classification of the dynein protein dataset utilization. The results and the 3080 best feature subsets from the BLOSUM62 technique are displayed in Table 3. In specific, based on BLOSUM62 profiles score with (XGBoost 0.91% ROC (AUC) curves score obtained. Dynein motor proteins that contain ROCs matching to one feature extraction approaches. Figure 4 illustrates how eXtreme gradient boosting (XGBoost) ROC-auc projected outputs performed better than those of other machine learning models.

3.5 Comparison metric performance of various machine learning classifiers

The BLOSUM62 matrix profiles score is a component of the results analysis. The BLOSUM62 matrix profiles with extreme gradient boosting (XGBoost) obtained accuracy score better than other four classification models—AdaBoost and RF, KNN,

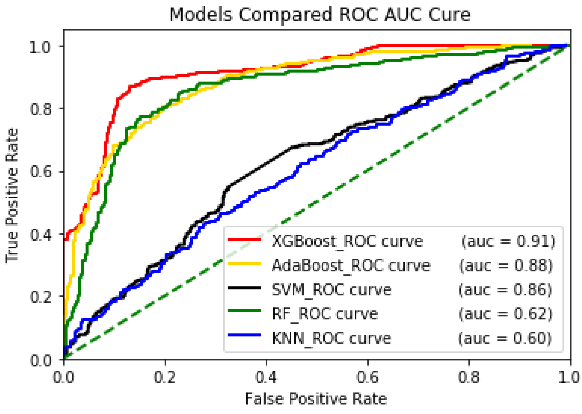

Figure. 4. Proposed model ROC-AUC score result

and SVM—were each regarded to have significant results. The BLOSUM62 model then recorded a precision score, indicating that the overall performance of our test's eXtreme gradient boosting (XGBoost) classification is good. Second, based on our experimental testing, the KNN classification was obtained, along with the highest analyses, as shown in Figure 5, and the third-best classifier with precision on the SVM model.

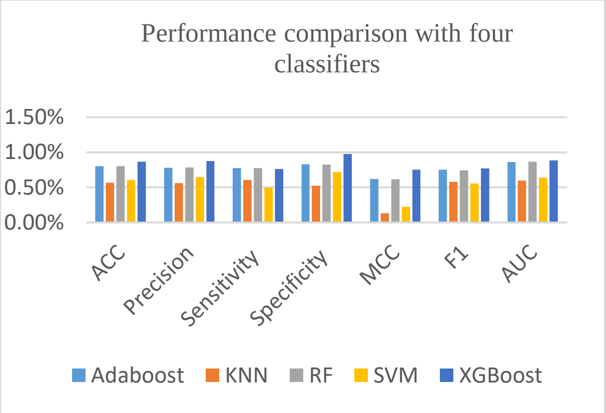

Figure. 5. Parameter metric performance evaluation of dynein protein.

CONCLUSION

An interesting side finding was that machine learning integrating with computational biology is a major concern for biological researchers in light of its outstanding results in a variety of fields. In this study, we introduced dynein-XGB, a predictor model built on the XGBoost algorithm for precise identification of dynein proteins. Specifically, when compared to earlier predictors on the benchmark dataset, we have attained state-of-the-art performance. There are three key inferences that can be made. First off, as compared to other algorithms, the XGBoost method performs dynein prediction with a higher level of stability and accuracy. Second, feature vectors were optimized using the feature selection technique known as Relief, which helped to extract key features from a wide pool of candidate features and enhance the model's functionality. Additionally, dynein-XGB, in contrast to previous sequence-based dynein predictors, can offer relevant explanation based on samples supplied utilizing the feature importance and the SHAP technique. When compared to conventional machine learning methods, our methodology improved on the majority of the analyzed metrics. In this paper, we established a reliable approach for accurately identifying novel proteins that are members of motor superfamilies, which can be exploited to develop therapeutic targets. The contributions of this study may serve as a foundation for future research that might tackle numerous bioinformatics issues.

DECLARATIONS

Funding: No funding was received for this study.

Conflicts of interest/Competing interests: The authors declare no any conflict of interest/competing interests.

Data availability: Not applicable

Code availability: Not applicable

Authors' contributions: Ali Ghulam and Rahu Sikander designed the concepts write-up, Dhani Bux Talpur, Sajjad Hussain Talpur and Erum Saba carried out the experiments and analyses and produced the manuscript. The method's development and the manuscript's revision were helped by Ali Ghulam. Zulfikar Ahmed Maher and Saima Tunio oversaw the project, offered helpful advice on how to execute it better, and revised the document. The article was read and approved by all writers.

REFERENCES

[1] S.A. Burgess, M.L. Walker, H. Sakakibara, P.J. Knight, K. Oiwa, Dynein structure and power stroke, Nature 421 (2003) 715–718.

[2] R.D. Vale, T.S. Reese, M.P. Sheetz, Identification of a novel force-generating protein, kinesin, involved in microtubule-based motility, Cell 42 (1985) 39–50.

[3] A.J. Roberts, T. Kon, P.J. Knight, K. Sutoh, S.A. Burgess, Functions and mechanics of dynein motor proteins, Nat. Rev. Mol. Cell Biol. 14 (2013) 713–726

[4] Hirokawa, N., Noda, Y., Tanaka, Y., & Niwa, S., Kinesin superfamily motor proteins and intracellular transport. Nature reviews Molecular cell biology, 10(10), (2009) 682-696.

[5] Banci, L., Bertini, I., Boca, M., Calderone, V., Cantini, F., Girotto, S., & Vieru, M., Structural and dynamic aspects related to oligomerization of apo SOD1 and its mutants. Proceedings of the National Academy of Sciences, 106(17), (2009) 6980-6985.

[6] Chen, X. J., Xu, H., Cooper, H. M., & Liu, Y., Cytoplasmic dynein: a key player in neurodegenerative and neurodevelopmental diseases. Science China Life Sciences, 57(4), (2014) 372-377.

[7] Eschbach, J., & Dupuis, L., Cytoplasmic dynein in neurodegeneration. Pharmacology & therapeutics, 130(3) (2011) 348-363.

[8] Bar-Or, A., Fawaz, L., Fan, B., Darlington, P. J., Rieger, A., Ghorayeb, C., ... & Smith, C. H., Abnormal B-cell cytokine responses a trigger of T-cell–mediated disease in MS?. Annals of neurology, 67(4) (2010) 452-461.

[9] Le, N. Q. K., Yapp, E. K. Y., Ou, Y. Y., & Yeh, H. Y., iMotor-CNN: Identifying molecular functions of cytoskeleton motor proteins using 2D convolutional neural network via Chou's 5-step rule. Analytical biochemistry, 575 (2019) 17-26.

[10] Zhu, C., Zhao, J., Bibikova, M., Leverson, J. D., Bossy-Wetzel, E., Fan, J. B., ... & Jiang, W., Functional analysis of human microtubule-based motor proteins, the kinesins and dyneins, in mitosis/cytokinesis using RNA interference. Molecular biology of the cell, 16(7) (2005) 3187-3199.

[11] Janssens, F., Glänzel, W., & De Moor, B., Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining (2007) (360-369).

[12] H. Khataee, A.W.-C. Liew, A mathematical model describing the mechanical kinetics of kinesin stepping, Bioinformatics 30 (2013) 353–359

[13] Dutta, M., & Jana, B., Computational modeling of dynein motor proteins at work. Chemical Communications, 57(3), (2021) 272-283.

[14] Li, L., Alper, J., & Alexov, E. (2016). Cytoplasmic dynein binding, run length, and velocity are guided by long-range electrostatic interactions. Scientific reports, 6(1), (2012)1-12.

[15] Erdős, G., Szaniszló, T., Pajkos, M., Hajdu-Soltész, B., Kiss, B., Pál, G., ... & Dosztányi, Z., Novel linear motif filtering protocol reveals the role of the LC8 dynein light chain in the Hippo pathway. PLoS computational biology, 13(12), (2017) e1005885.

[16] Gao, F. J., Hebbar, S., Gao, X. A., Alexander, M., Pandey, J. P., Walla, M. D., ... & Smith, D. S., GSK-3β phosphorylation of cytoplasmic dynein reduces Ndel1 binding to intermediate chains and alters dynein motility. Traffic, 16(9), (2015) 941-961.

[17] Ho, Q. T., & Ou, Y. Y., Classifying the molecular functions of Rab GTPases in membrane trafficking using deep convolutional neural networks. Analytical biochemistry, 555, (2018) 33-41.

[18] Zou, C., Gong, J., & Li, H., An improved sequence-based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis. BMC bioinformatics, 14(1), (2013)1-14.

[19] Zou, Q., Wan, S., Ju, Y., Tang, J., & Zeng, X., Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. BMC systems biology, 10(4), (2016)401-412.

[20] Tao, Z., Li, Y., Teng, Z., & Zhao, Y., A method for identifying vesicle transport proteins based on LibSVM and MRMD. Computational and Mathematical Methods in Medicine, (2020).

[21] Kumar, K., & Thakur, G. S. M., Advanced applications of neural networks and artificial intelligence: A review. International journal of information technology and computer science, 4(6), (2012) 57.

[22] Zhang, Y., Qiao, S., Ji, S., Han, N., Liu, D., & Zhou, J., Identification of DNA–protein binding sites by bootstrap multiple convolutional neural networks on sequence information. Engineering Applications of Artificial Intelligence, 79, (2019)58-66.

[23] Arif, Muhammad, et al. "StackACPred: prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach." Chemometrics and Intelligent Laboratory Systems 220 (2022): 104458.

[24] Arif, Muhammad, et al. "DeepCPPred: a deep learning framework for the discrimination of cell-penetrating peptides and their uptake efficiencies." IEEE/ACM Transactions on Computational Biology and Bioinformatics 19.5 (2021): 2749-2759.

[25] Ge, Fang, et al. "TargetMM: Accurate Missense Mutation Prediction by Utilizing Local and Global Sequence Information with Classifier Ensemble." Combinatorial Chemistry & High Throughput Screening 25.1 (2022): 38-52.

[26] Ghulam, Ali, et al. "Accurate prediction of immunoglobulin proteins using machine learning model." Informatics in Medicine Unlocked 29 (2022): 100885.

[27] Ghulam, Ali, et al. "ACP-2DCNN: deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network." Chemometrics and Intelligent Laboratory Systems 226 (2022): 104589.

[28] Ghulam, Ali, et al. "Disease-pathway association prediction based on random walks with restart and PageRank." IEEE Access 8 (2020): 72021-72038.

[29] J. Song, F. Li, K. Takemoto, G. Haffari, T. Akutsu, K.-C. Chou, G.I. Webb, PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework, J. Theor. Biol. 443 (2018) 125–137.

[30] G.O. Consortium, Expansion of the gene Ontology knowledgebase and resources, Nucleic Acids Res. 45 (2016) D331–D338.

[31] Jia, K., & Jernigan, R. L., New amino acid substitution matrix brings sequence alignments into agreement with structure matches. Proteins: Structure, Function, and Bioinformatics, 89(6), (2021)671-682.

[32] Sakhanenko NA, Galas DJ. Biological data analysis as an information theory problem: multivariable dependence measures and the shadows algorithm. J Comput Biol. 2015;22:1005-1024.

[33] Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS ONE 2017, 12, e0177678.

[34] Ding, Y.; Tang, J.; Guo, F. Identification of drug–target interactions via fuzzy bipartite local model. Neural Comput. Appl. 2020, 32, 1–17.

[35] Lee, B.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol., 1971, 55, 379-400.

[36] Statnikov, A.; Wang, L.; Aliferis, C.F. A Comprehensive comparison of random forests and support vector machines for microarray-based cancer classification, BMC Bioinformatics, 2008, 9, 319.

[37] Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000; 16:412–424.

[38] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975; 405:442–451.

[39] Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging–SVM ensemble classifier. Artif. Intell. Med. 2019, 98, 35–47.

[40] Jiang, Q.; Wang, G.; Jin, S.; Yu, L.; Wang, Y. Predicting human microRNA–disease associations based on support vector machine. Int. J. Data Min. Bioinform. 2013, 8, 282–293.

[41] Murugan, A.; Nair, S.A.H.; Kumar, K.P.S. Detection of Skin Cancer Using SVM, Random Forest and KNN Classifiers. J. Med. Syst. 2019, 43, 269.

Received: 15 Sep. 2022. Revised/Accepted: 10 Nov. 2022.