

# Decision Tree Based ML-KNN Approach For Protein Sub-Cellular Localization Prediction

**Komal Arora,**

Assistant Professor,

School of Computer Science Engineering, Lovely Professional University, Punjab (India),

[Komal.17783@lpu.co.in](mailto:Komal.17783@lpu.co.in)

**Jaspreet Kaur,**

Research Scholar

School of Computer Science Engineering, Lovely Professional University, Punjab (India),

[jaspreetk200@gmail.com](mailto:jaspreetk200@gmail.com)

**Abstract-** *Sub-cellular localization of different Protein is an important study on proteomics, geonomics, cytobiology and drug analysis etc. There are so many fields of biology where the use of computational methods results in drastic advancements in biomedical sciences and results in great benefits to medical field. Predicting sub-cellular localization of different proteins is significant to get the organic capacities, functions and processes just as finding new medications. This paper proposed a technique ML-KNN with decision tree, which is obtained after enhancing KNN algorithm. Various parameters are used to evaluate the final outputs. Experimental results after this study represents how proposed system is gives better performance than that of the current procedure.*

**Keywords** –protein sub-cellular localization, biology, proteins, ml-knn, decision tree.

## I. INTRODUCTION

Cell is the essential unit of life and proteins are the workhorses in the cell. Aside from few proteins, coded in the genomes of mitochondria and chloroplasts, the guidelines for making the vast majority of the proteins in a cell are composed in protein encoding qualities of DNA, which dwells in the core of the cell. At the season of quality articulation, this data is deciphered into mRNA (emissary RNA), and removed from core to the cytoplasm. With the assistance of tRNAs (exchange RNA), the mRNA is converted into protein at the ribosome exhibit in the cytoplasm. Once the protein is made, it ought to be translocated to its objective organelle to play out its planned capacity. Proteins are arranged to its sub cellular area with the assistance of an inalienable flag show in the essential grouping level. In 1999, Gunter Blobel was granted Nobel Prize in Physiology

or Medicine for the disclosure that "proteins have natural flags that oversee their vehicle and limitation in the cell". These address signals are normally present in the N-terminal of the amino corrosive grouping.

To understand the biological processes at the cellular level, in this era researchers are preferring to study proteomics inside out.. One of the significant goal in proteomics is protein sub-cellular localization and that has received a lot of attention recently because it is the main attribute of a protein. Information about where a protein resides in a cell is important to determine how they interact with each other and with other molecules and knowing this information let us know about not only the function of proteins but also the organization of the whole cell [1]. However, it is time-consuming, expensive, and laborious process.

Nowadays, discovery of various protein results in new types of proteins, and it makes traditional techniques more unpractical to implement [3]. Recent advancements in proteomics and genomics methods bought a scientific revolution and resulted in large accumulation of proteins whose functions are unknown. Due to this advancement, many powerful tools for predicting the sub-cellular location of proteins have been proposed which leads to the proposal of various new methods so that to predict the protein sub-cellular localization [2]. There is a need to use an algorithm, which can identify multi sites subcellular localization fast and reliable.

## II. BACKGROUND

Xumi et al. [4] In this paper, a few component extraction techniques were melded to extricate the element data, at that point the multi-mark k closest neighbors (ML-KNN) calculation was utilized to foresee protein sub-cell areas. The best general exactnesses we got for dataset s1 in developing Gpos-mploc is 66.7304 and 59.9206 percent for dataset s2 in constructing Virus-mPLoc.

Wang et al. [6] proposed a novel method Random Label Selection (RALS) derived from multi-label learning via RALS, which expands the simple Binary Relevance (BR) technique. RALS endeavors to take in the mark connections from information by enlarging unique element space with arbitrarily chosen names as its extra information highlights. Test comes about on two benchmark informational indexes likewise demonstrate that our proposed strategies accomplish altogether higher execution than some other best in class techniques in anticipating subcellular multi-locations of proteins.

Joseph et al. [9] In this paper, the author describes the discriminative motif finding which is used to predict protein sub-cellular localization and then also fused discriminative motif finding so that to sort protein with hierarchical structure. With this approach the author, find the potential errors in public databases which are there to find where the protein resides in a cell..

Shen et al. [10] In this paper, the author, uses Euk-mPLoc 2.0 tool for the prediction of the location of protein in a cell with in single as well as in multiple sites. The tool Euk-mPLoc is

freely available on the internet and it is a novel approach to get highly accurate results.

Chou et al. [11] In this paper, the author, finds the protein sub-cellular location in multiple sites using a fusion classifier to find viral proteins. The protein sub-cellular location found using the Virus-mPLoc tool.

Juan et al. [14] utilized various protein portrayal strategies and three regular classifiers. Expansive scale protein examination and dependable comments are especially useful for researchers in science and prescription network. Understanding the practical portrayals of protein successions has been a noteworthy test lately. Broad PC based forecast frameworks have been produced to help the need since numerous proteins' subcellular confinements are as yet obscure. These protein depiction techniques are portrayed into two classes: protein structure-based and position-explicit scoring network based. A prevalent measure is rehearsed upon broadly utilized illuminating arrangements. Through these assessments, it is relied on to give an evaluation between protein delineation strategies for protein subcellular obstruction and their depiction qualities.

Wan et al. [15] proposes joining homology-based profile arrangement techniques and utilitarian area based Gene Ontology (GO) strategies to anticipate the subcellular areas of proteins. The elements of proteins are firmly identified with their subcellular areas. Computational techniques are required to supplant the relentless and tedious trial forms for proteomics examine. The component vectors built by these two strategies are perceived by

help vector machine (SVM) classifiers, and their scores are intertwined to upgrade arrangement execution. The paper additionally examines distinctive ways to deal with building the GO vectors in light of the GO expressions came back from InterProScan. The outcomes exhibit that the GO strategies are equivalent to profile-arrangement techniques and eclipse those in light of amino-corrosive pieces. Likewise, the combination of these two strategies can beat the individual techniques.

### III. PROPOSED WORK

Proposed technique works as follows:

1. Conventional Decision Tree algorithm will be applied on proteins with single site and on the proteins with multiple sites inputted into algorithm.
2. The ML-KNN algorithm is fused with decision Tree to predict the protein sub-cellular locations. The training of ML-KNN is optimized using Decision Tree. This optimized training is further used to predict the protein sequences.

3. The accuracy can be checked by the equation,

$$\text{Accuracy (x)} = 100 - \text{error rate}$$

Where, x can be any sub-cellular location.

4. Analyse the performance parameters like, prediction accuracy, prediction time and efficiency and then compare the

results with existing methodologies for protein location prediction.

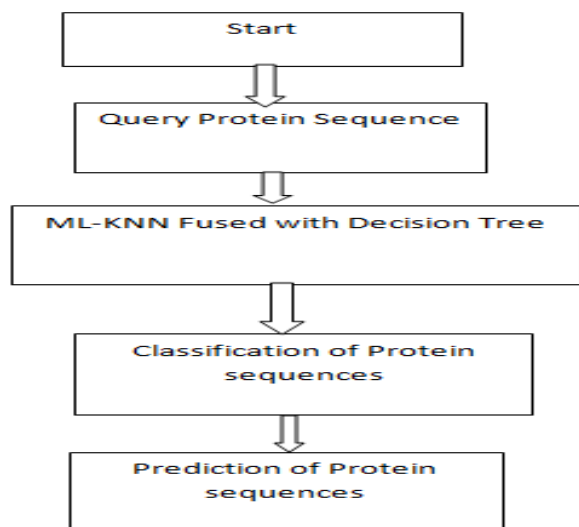


Figure 1: Flowchart of Proposed Technique

KNN - K-Nearest Neighbors (KNN) is one of the least difficult calculations utilized as a part of Machine Learning. KNN is a non parametric lazy learning calculation. KNN calculations utilize information and group new information focuses in view of a similitude measures (e.g. remove work). Classification is finished by a dominant part vote to its neighbors. The information is allocated to the class which has the closest neighbors. As you increment the quantity of closest neighbors, the estimation of k, accuracy may increment.

ML-KNN - Machine Learning is a standout amongst the most well known methodologies in Artificial Intelligence. Over the previous decade, Machine Learning has turned out to be one of the basic parts of our life. It is actualized in an undertaking as straightforward as perceiving human penmanship or as difficult as self-driving autos. Multi-label learning in began

from the examination of content classification issue, where each record may have a place with a few predefined themes at the same time. In multi-label learning in, the planning set is made out of models each related with a course of action of names, and the endeavor is to envision the name sets of unnoticeable cases through separating getting ready events with realized name sets. In this examination work a calculation named ML-KNN with decision tree is displayed, which is gotten from the conventional k-Nearest Neighbor (kNN) calculation.

DT – A graph which uses branching method to to give us every possible outcome and help us in deciding to most appropriate result is decision Tree. It is one way to deal with demonstrate a computation that solitary contains unexpected control articulations. Decision trees are typically used as a piece of exercises analyse, especially in choice examination, to help perceive a framework bound to accomplish a target, however in the meantime are a standard instrument in machine learning..

ML-KNN with DT – Previous work in the domain of protein sub-cellular localization is only based on KNN. To optimize the results of existing technique, the ML-KNN algorithm is fused with decision Tree to predict the protein sub-cellular locations. The training of ML-KNN is optimized using Decision Tree. This optimized training is further used to predict the protein sequences. This work is expected to result with an algorithm, which results into the better performance with improved accuracy of existing methodology.

#### IV. EXPERIMENTAL RESULTS

This section presents experimental results of the proposed technique.

Proposed technique, ML-KNN with DT is evaluated on various parameters and is compared with KNN. Performance of proposed technique is evaluated on basis of various parameters:

- Accuracy
- Precision
- Execution Time
- Error Rate



Figure 2: Showing results of Proposed Technique

Above figure shows results of the proposed technique. Precision of proposed technique is 0.0614, error rate of proposed technique is 13.8560, accuracy of the proposed technique is 86.1440, and execution time of the proposed technique is 51.2196.

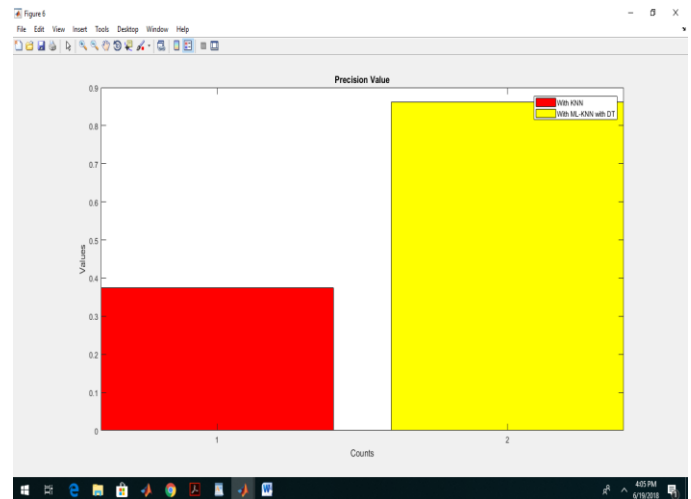


Figure 3: Showing comparison of KNN with ML-KNN DT on the basis of precision

Figure 3 shows comparison of KNN with ML-KNN DT on the basis of precision. It is clear from the figure that precision of ML-KNN with DT is more than KNN.

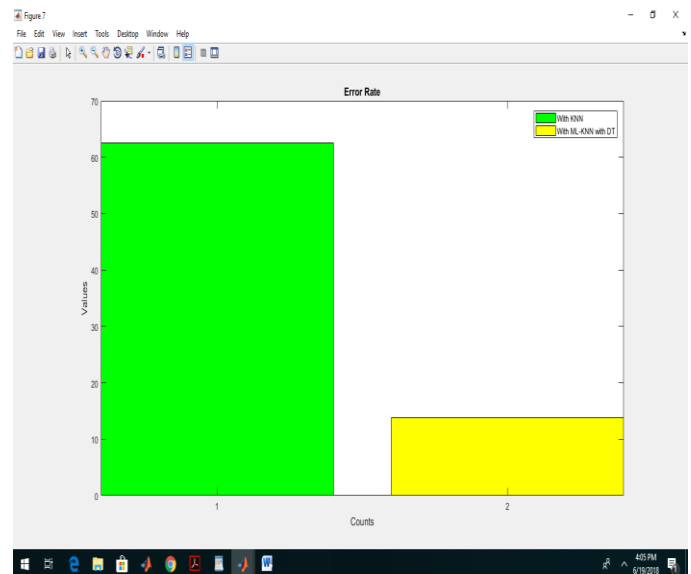


Figure 4: Showing comparison of KNN with ML-KNN DT on the basis of error rate

Figure 4 shows comparison of KNN with ML-KNN DT on the basis of error rate. It is clear from the figure that error rate of ML-KNN with DT is less than KNN.

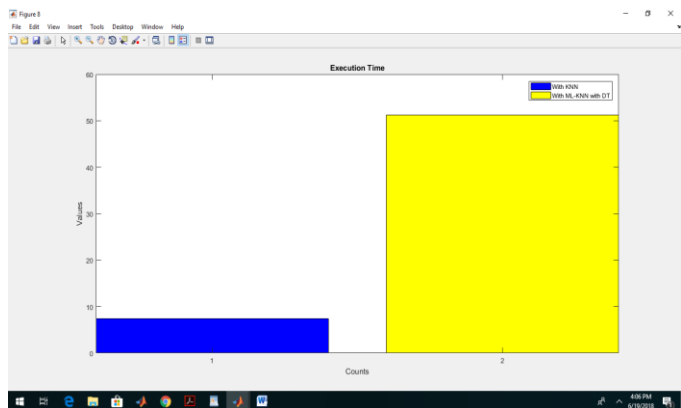


Figure 5: Showing comparison of KNN with ML-KNN DT on the basis of execution time

Figure 5 shows comparison of KNN with ML-KNN DT on the basis of execution time. It is clear from the figure that execution time of ML-KNN with DT is more than KNN.

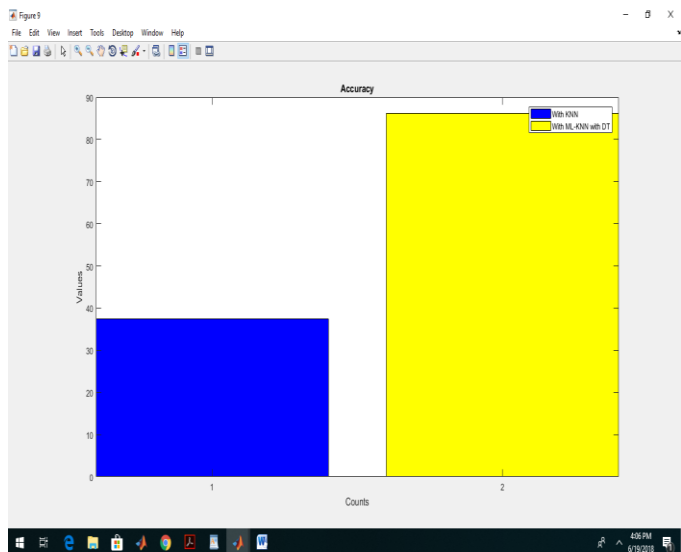


Figure 6: Showing comparison of KNN with ML-KNN DT on the basis of accuracy

Figure 6 shows comparison of KNN with ML-KNN DT on the basis of accuracy. It is clear from the figure that accuracy of ML-KNN with DT is more than KNN.

## V. CONCLUSION

The primary goal of this investigation is to discover a strategy for proficient subcellular restriction with more exactness in lesser time. The Decision Tree (DT) classification algorithm can be used to classify proteins sub-cellular locations and its fusion with threshold value will result in improved accuracy and better performance. Prediction of multi-sites protein sub-cellular localization would help to study biology in a better way.

Proposed technique is implemented using MATLAB. Various parameters are used to check and get the performance of the proposal such as accuracy, precision, error rate and execution time. Results of the proposed technique (ML-KNN with DT) are compared with the existing technique (KNN). Experimental results demonstrate that accuracy, precision and execution time of the proposed technique (ML-KNN with DT) is more than the existing technique (KNN) and error rate of the proposed technique (ML-KNN with DT) is less than the existing technique (KNN).

In this work we use protein sequences, in future instead of using protein sequences we may use any other dataset of DNA sequences. Also results we got after certain iterations shows that the time taken by ML-KNN with DT to execute is less than the previous technique, so to improve it further, instead of using decision tree we may embed any other algorithm for

optimization so that execution time will decrease and we get more enhanced results.

## REFERENCES

- [1] U. Subhashini, P. Bhargavi, S. Jyothi, D.M. Mamatha, "Predicting Sub-cellular Localization of Proteins with Multiple Sites using Threshold ML-KNN", *Int J Pharma Bio Sci* 2017 July; 8(3): (B) 278-285
- [2] Shalini Kaushik, Usha Chouhan, Ashok Diwedi "Study of Protein Sub-cellular Localization Prediction-A Review", *International Journal of Life Science and Pharma Research*, Vol 7/Issue 3/ 2017, ISSN 2250-0480
- [3] Kirtan Dave, Nikita Patel, Dr. Hetal Panchal, "Prediction of Sub-cellular Localization of Human Protein by Support Vector Machine", under consideration for Young Scientist of India award.
- [4] Xumi Qu, Dong Wang, Yuehui Chen, Shanping Qiao, Qing Zhao, "Predicting the Subcellular Localization of Proteins with Multiple Sites Based on Multiple Features Fusion", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol 13, No. 1, January/February 2016.
- [5] Xiao Wang, Guo-Zheng Li, "Multi-label Learning via Random Label Selection for Protein Sub-cellular Multi-locations Prediction", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol 10, No. 2, March/April 2013.
- [6] Tien-ho Lin, Robert F. Murphy and Ziv Bar-Joseph, "Discriminative Motif Finding for Predicting Protein Subcellular Localization", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol 8, No. 2, March-April, 2011.
- [7] Kuo-Chen Chou, Hong-Bin Shen, "A New Method for Predicting the Subcellular Localization of Eukaryotic Proteins with Both Single and Multiple Sites: Euk-mPLoc 2.0", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, February, 2010.
- [8] Hong-Bin Shen and Kuo-Chen Chou, "Virus-mPLoc: A Fusion Classifier for Viral Protein Subcellular Location Prediction by Incorporating Multiple Sites", *Journal of Biomolecular Structure & Dynamics*, ISSN 0739-1102 Volume 28, Issue Number 2, (2010).
- [9] Min-Ling Zhang and Zhi-Hua Zhou, "A Review on Multi-Label Learning Algorithms", *IEEE/ACM Transactions on Computational Biology* September, 2007.
- [10] Eric Y.T. Juan, J.H. Chang, C.H. Li, B.Y. Chen, "Methods for Protein Subcellular Localization Prediction", *IEEE, International Conference on Complex, Intelligent, and Software Intensive Systems*, 2011, pp. 553-558.
- [11] Shibiao Wan, Man-Wai Mak, Sun-Yuan Kung, "Protein Subcellular Localization Prediction Based on Profile Alignment and Gene Ontology", *IEEE, International Workshop on Machine Learning for Signal Processing*, 2011.