NIPES

Journal of Science and Technology Research

Journal homepage: www.nipesjournals.org.ng

# Machine Learning Algorithm for Predicting Subcellular Localization Sites for Proteins

***Osaseri R.O[1] & Usiobaifo A.R.[1]***

[1]Department of Computer Science, Faculty of Physical Sciences, University of Benin, P.M.B. 1154, Benin City, Nigeria

Email: **roseline.osaseri@uniben.edu, rosemary.usiobaifo@uniben.edu**

## Article Info

***Keywords:*** *Subcellular Localization sites, Protein, machine learning, KNN, logistic regression*

*Received 12 August 2024*
*Revised 3 November 2024*
*Accepted 4 November 2024*
*Available online 8 December 2024*

https://doi.org/10.5281/zenodo.14302877



## Abstract

*Predicting the location of a protein within the cell can help in elucidating its function and deducing its involvement in certain biochemical pathways. In this study, machine learning models are investigated to predict the Protein subcellular Localization Sites in Cells. The aim of this research is to develop an algorithm that can accurately predict the Protein subcellular Localization Sites in Cells Processes that help in determine healthy cell which is crucial for understanding protein functions and their roles in various biochemical pathways, the onset of disease and its potential use as a drug target. Two models were explored, which are Logistic Regression; A statistical model suitable for binary classification, which estimates probabilities of localization based on input features and K-Nearest Neighbor (KNN); a non-parametric method that classifies proteins based on the majority label of their nearest neighbors in the feature space for prediction of the Protein subcellular Localization Sites in Cells. A comprehensive dataset containing protein sequences and their corresponding subcellular localization labels was curated. Relevant features from the protein sequences were extracted; the dataset was divided into training and testing sets. Models were trained on the training set, and their performance was evaluated on the testing set. Model performance was assessed using several key metrics: Classification Accuracy, F-score Precision and Recall which was found through confusion matrix. K-Nearest Neighbors (KNN) achieved the highest accuracy of 98% and a precision of 100%, indicating it correctly classified almost all instances and did not misclassify any positives. Logistic Regression demonstrated a classification accuracy of 92%, with precision and recall values of 96%. While it performed well, it was not as effective as KNN in this context. The confusion matrix provided insights into the model performance, revealing rates of true and false positives, which are crucial for understanding the models' strengths and weaknesses. The findings suggest that K-Nearest Neighbors (KNN) is the more suitable model for predicting protein subcellular localization sites in cells, offering higher accuracy and precision compared to Logistic Regression.*

## 1.0. Introduction

A cell is the most basic unit of life, and all living things are made of cells, cells arise from the [1]. Over the years, cell biology has progressed in steps to understand and characterize cells. There are different types of cells, despite this difference, the cell organization is similar, and they share

common features, including that they all self-replicate and are separated from the extracellular space by the cell membrane, which allows substances to go in and out of the cell. The cell membrane is present in every type of cell.

The biological cell is a complex structural unit with various functionally distinct subcellular compartments/ locations. These subcellular compartments include the cell membrane, cytoplasm, nucleus, endoplasmic reticulum, Golgi apparatus, mitochondria, and extracellular region, each with a defined set of roles. The major role of subcellular localization is to provide a functional environment for proteins [2].

Many experiments have been developed for accurately recognizing subcellular locations ofproteins [3]. However, wet experiments are characterized by long experimental time, high experimental failure rate, and expensive experimental materials. In order to avoid these disadvantages, machine learning-based methods were developed to predict protein subcellular location [4],[5]. Computational approaches are becoming indispensable components of molecular and cellular biology, especially in the analyses of complex genomes for which massive amounts of sequence data must be examined for biological function. Functional information can be obtained from sequence information not by solving equations of first principles, but by inference based on empirical knowledge. Although the sequences data are now collected and organized in publicly available database, functional data are not well organized, except, perhaps, in the brain of a human expert.

Protein sequencing played a pivotal role in mapping out the human genome, and is an essential tool for many basic and applied research applications today. It is an important tool for determiningthe thousands of nucleotide variations associated with specific genetic diseases, like Huntington's,which may help to better understand these diseases and advance treatment. The accurate identification of subcellular localization of a protein is a crucial step for its functional annotation and to decide its role in underlying complex biological processes.

Experimental techniques to characterize proteins at a biochemical, structural and physiological level have improved considerably over the years providing researchers with the tools necessary to understand protein function at a cellular and an organism level. Combined with detailed functional data, large-scale genome sequencing efforts have also greatly increased the scale of proteomic data available from model and non-model species. However, a major challenge facing researchers todayis simply keeping pace with the sheer volume of low throughput and high-throughput data being generated. Although scientific publications are used to disseminate research findings to the wider community, manually identifying, curating and collating individual experiments is a time and labour intense process. Understanding protein subcellular localization is important to help understand not only the function of individual proteins but also the organization of the cell as a whole. The traditional approach to determine the subcellular localization of protein depends on biochemical experimentssuch as fluorescence microscopy, electronic microscopy, and cell separation methods [6]. However, for a single protein, these methods are very labor-intensive and often time-consuming in today’s post-genomic era, given the rate at which protein data is generated. A reliable automated method is required that can precisely predict the subcellular localization of protein molecules [7]. Automating this process with higher accuracy remains a challenging task in molecular/computational biology. In this study we have made an attempt to develop an approach that analyzes machine learning algorithm in Predicting Cellular Protein localization Sites on E-coli’s dataset.

### 1.1 Review of Related Work

[8] did a comparison of four classifiers to predict cellular localization sites of proteins in yeast and E.coli. A set of sequence derived features, such as regions of high hydrophobicity, were used for each classifier. The methods compared were a structured probabilistic model specifically designed for the localization problem, the k nearest neighbors classier, the binary decision tree classifier, and the naive Bayes classifier. The result of tests using stratified cross validation show that k nearest neighbor classifier performs better than other methods. In the case of yeast this difference was statistically significant using a cross-validated paired t test. The result is an accuracy of approximately 60% for 10 yeast classes and 86% for 8 E.coli classes. The best previously reported accuracies for these datasets were 55% and 81% respectively.

[9] Investigated a meta-learning approach for classifying proteins into their various cellular locations based on their amino acid sequences. A meta-learner system based on k-Nearest

Neighbors (k-NN) algorithm as base-classifier, since it has shown good performance in this context as individual classifier and DECORATE as meta-classifier using cross-validation tests for classifying Escherichia Coli bacteria proteins from the amino acid sequence information is evaluated. A report of comparison against a Decision Tree induction as base-classifier is also evaluated. The experimental results show that the k-NN-based meta-learning model is more efficient than the Decision Tree-based model and the individual k-NN classifier. Results of KNN gave 87.5% accuracy obtained using 5- CV on E.coli dataset. Its Confusion Matrix also shows that none of the minority class proteins namely imL and imS, have been classified correctly.

[10] present a Support Vector Machines- Recursive Feature Elimination (SVM-RFE) Feature selection technique to select suitable features from the many features in the Bakers Yeast dataset. The obtained features were used for predicting essential proteins. The goal of feature selection was to find the suitable features that both have powerful prediction ability for protein essentiality and share minimal biological meaning between each other. The SVM-RFE algorithm adopts a backward feature elimination strategy. It constructs sorting coefficient by weight vectors generated by Support Vector Machine (SVM), and then removes iteratively a feature with the smallest coefficient. SVM-RFE gets the sorted list in descending order of all the features.

[11] Presents a backward feature selection technique that is applied to thousands of features on three datasets including M638 which contains 638 proteins, Gneg1456 including 1456 locative proteins and Gpos523 consisting of 523 Gram-positive bacterial protein sequences within each subcellular localization. Backward feature selection technique is used here to rank the features so as to find out the informative features and reduce the computation cost. The initial feature vector for each protein is constructed by combining PSSM, PROFEAT and GO features. For each dataset, feature vectors of all proteins constituted a feature matrix, where each row corresponded to a sample and each column corresponded to a feature. Then, SVM-RFE is implemented by training an SVM with a linear kernel on the feature matrix. The top K features are finally obtained by eliminating a number of features corresponding to the smallest ranking criteria and applied in sequel.

[12] Proposed a feature subset selection technique whereby the statistical significance of each feature of a superfamily from all other superfamilies is measured. This technique was applied on a protein sequence represented by a vector of 8420 features. The features that did not contribute in the representation of a sequence were removed from the original feature space to substantially reduce feature vectors' dimension. The proposed feature selection technique extracts different

subsets of features from the original feature space and selects the best feature subset that shows maximum accuracy results. The subset of the best and relevant features was used to discriminate between different protein classes or superfamilies. The processed data, after the feature selection, is used during the classification which drastically minimizes the running time of the Classification algorithms.

Constituent amino-acids can be analyzed to predict secondary, tertiary and quaternary protein structure. Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence. That is, the prediction of its folding and its secondary, tertiary, and quaternary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes).

There are three major theoretical methods for predicting the structure of proteins: Comparative modeling, Fold recognition and Abinitio prediction.

### 1) Comparative Modeling

Comparative modeling exploits the fact that evolutionarily related proteins with similar sequences, as measured by the percentage of identical residues at each position based on an optimal structural superposition, have similar structures. The similarity of structures is very high in the so-called

``core regions", which typically are comprised of a framework of secondary structure elements such as alpha-helices and beta-sheets. Loop regions connect these secondary structures and generally vary even in pairs of homologous structures with a high degree of sequence similarity.

### 2) Fold Recognition

Threading uses a database of known three-dimensional structures to match sequences without known structure with protein folds. This is accomplished by the aid of a scoring function that assesses the fit of a sequence to a given fold. These functions are usually derived from a database of known structures and generally include a pairwise atom contact and solvation terms. Threading methods compare a target sequence against a library of structural templates, producing a list of scores. The scores are then ranked and the fold with the best score is assumed to be the one adopted by the sequence. The methods to fit a sequence against a library of folds can be extremely elaborate computationally, such as those involving double dynamic programming, dynamic programming with frozen approximation, Gibbs Sampling using a database of ``threading" cores, and branch and

bound heuristics, or as “simple” as using sophisticated sequence alignment methods such as Hidden Markov Models.

### 3) Abinitio Prediction

The abinitio approach is a mixture of science and engineering. The science is in understanding how the three-dimensional structure of proteins is attained. The engineering portion is in deducing the three-dimensional structure given the sequence. The biggest challenge with regards to the folding problem is with regards to abinitio prediction, which can be broken down into two components: devising a scoring function that can distinguish between correct (native or native-like) structures from incorrect (non-native) ones, and a search method to explore the conformational space. In many abinitio methods, the two components are coupled together such that a search function drives, and is driven by, the scoring function to find native-like structures.

Assigning subcellular localization to proteins is one of the major tasks of functional proteomics. Despite the impressive technical advances of the past decades, it is still time-consuming and laborious to experimentally determine subcellular localization on a high throughput scale. Thus, computational predictions are the preferred method for large-scale assignment of protein subcellular localization, and if appropriate. Previous studies indicated that proteins in the same organelle share specific functional domains [13].

Following extraction of a training dataset containing proteins clearly assigned to one of the 14 sub-locations, we performed the prediction work in a feature space constructed using the protein domain composition, as obtained from a well-established database using machine learning approach.

## 2.0. Methodology

This study employed supervised machine learning approach in order to mitigate the limitation of the existing system. Logistic regression and K-Nearest Neighbor prediction algorithms were usedto build the prediction models for protein subcellular localization sites.

A. Logistic regression

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. It can predict the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No,0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilisticvalues which lie between 0 and 1. Logistic Regression is much like Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereasLogistic regression is used for solving the classification problems.

B. K-Nearest Neighbor

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. The algorithm assumes similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. It stores all the available data and classifies a new data point based on the similarity. This means when new data

appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. It is a non-parametric algorithm that does not make any assumption on underlying data.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action onthe dataset. When the algorithm is at its training phase, its just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

### 2.1. Data Collection

The dataset for this study was collected from Kaggle machine learning repository; with datasets that having 336 instances of Ecoli protein sequences. These are classified into 8 different classes namely cp, im, imS, imL, imU, om, omL and pp. The dataset is extremely imbalanced because thedistribution of instances in each class is very much variant. The percent of representation of the eight classes is represented in Table 1.

Table 1: Protein sites representation.

| S/N | Site | % of the Dataset |
|---|---|---|
| 1 | cp | 42.56 |
| 2 | im | 22.92 |
| 3 | imS | 0.6 |
| 4 | imL | 0.6 |
| 5 | imU | 10.42 |
| 6 | om | 5.95 |
| 7 | omL | 1.49 |
| 8 | PP | 15.48 |

### 2.2.Data Preparation

Data preparation includes preprocessing such as fixing missing values and scaling of the variables into numeric. These processes prepare the data for modeling. There were no missing vales in the 336 datasets

### 2.3.Data Preprocessing

Counting number of unique classes of the dataset such as sites, sequence name and number of counts was carried out. The preprocessing of the dataset involved several critical steps to ensure its suitability for analysis. The first step was to count the number of unique classes within the dataset, which included the following categories:

a. Sites: Each unique site where the data was collected was identified and counted. This helps in understanding the geographical or experimental diversity within the dataset.
b. Sequence Names: The dataset was examined for unique sequence identifiers. This step is vital for tracking and comparing different sequences throughout the analysis.
c. Counts: The total number of occurrences for each unique class was recorded. This quantitative

assessment is essential for statistical analyses and understanding the distribution of data points. The results of this analysis are summarized in Table 2, which provides a comprehensive overview of the unique classes identified in the dataset. Each entry in the table includes the category, the number of unique entries, and additional relevant statistics that may assist in further analyses. Additionally, the distribution of these unique classes is visually represented in Figure 1. This figure includes distribution plots that illustrate the frequency of each class, enabling a quick assessment of the dataset's composition. Such visualizations are crucial for identifying any imbalances or anomalies in the data, which could influence subsequent analytical outcomes.

Table 2: Counting number of unique classes of the dataset

| | SEQUENCE_NAME | MCG | GVH | LIP | CHG | AAC | ALM1 | ALM2 |
|---|---|---|---|---|---|---|---|---|
| | count | count | count | count | count | count | count | count |
| SITE | | | | | | | | |
| cp | 143 | 143 | 143 | 143 | 143 | 143 | 143 | 143 |
| im | 77 | 77 | 77 | 77 | 77 | 77 | 77 | 77 |
| imL | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| imS | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| imU | 35 | 35 | 35 | 35 | 35 | 35 | 35 | 35 |
| om | 20 | 20 | 20 | 20 | 20 | 20 | 20 | 20 |
| omL | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| pp | 52 | 52 | 52 | 52 | 52 | 52 | 52 | 52 |

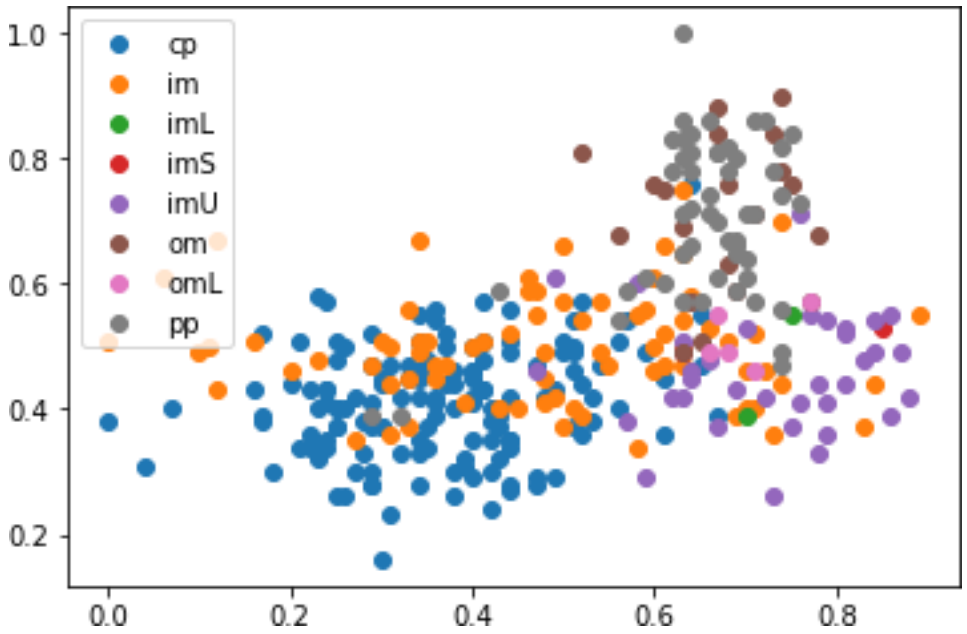


**Figure 1: Counting number of unique classes of the dataset the distribution plots**

The raw dataset were saved with an extension of CSV format in Microsoft excel 2010 version. The preprocessing process started by cleaning the raw data, the raw dataset consists of categorical values real numbers and integers. A sample of the raw dataset and the scaled data set are represented in table 3 and 4 respectively.

**Table 3: Input data head and tail**

| MCG | GVH | LIP | CHG | AAC | ALM1 | ALM2 | SITE |
|---|---|---|---|---|---|---|---|
| 270 | 0.56 | 0.68 | 0.48 | 0.5 | 0.77 | 0.36 | 5 |
| 120 | 0.25 | 0.26 | 0.48 | 0.5 | 0.39 | 0.32 | 0 |
| 259 | 0.78 | 0.68 | 0.48 | 0.5 | 0.83 | 0.4 | 5 |
| 233 | 0.66 | 0.48 | 0.48 | 0.5 | 0.54 | 0.7 | 4 |
| 167 | 0.47 | 0.59 | 0.48 | 0.5 | 0.52 | 0.76 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 323 | 0.76 | 0.73 | 0.48 | 0.5 | 0.44 | 0.39 | 7 |
| 192 | 0.41 | 0.51 | 0.48 | 0.5 | 0.53 | 0.75 | 1 |
| 117 | 0.51 | 0.49 | 0.48 | 0.5 | 0.53 | 0.14 | 0 |
| 47 | 0.43 | 0.4 | 0.48 | 0.5 | 0.39 | 0.28 | 0 |
| 172 | 0.33 | 0.45 | 0.48 | 0.5 | 0.45 | 0.88 | 1 |

**Table 4: Input data scaled**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -0.06819 | -0.12186 | -0.22361 | 0 | -2.57894 | -1.46458 | -1.02678 | -0.78344 |
| -1.33549 | -0.87754 | -0.22361 | 0 | -1.45426 | -0.39883 | -0.04134 | -0.78344 |
| 1.04703 | 0.702521 | -0.22361 | 0 | 2.179327 | -0.39883 | -1.26141 | 1.129609 |
| 0.844262 | 0.152935 | -0.22361 | 0 | 1.227674 | 0.296227 | -0.6983 | 1.129609 |
| -0.57511 | -0.12186 | -0.22361 | 0 | -1.1082 | 1.315643 | 1.460282 | -0.40083 |
| 1.401874 | 1.5269 | -0.22361 | 0 | -0.15655 | 1.083957 | 1.27258 | 0.747 |
| -1.18342 | -0.60275 | -0.22361 | 0 | -0.5026 | -0.16714 | 0.193289 | -0.78344 |
| 0.742878 | 2.145184 | -0.22361 | 0 | -0.5026 | -0.76953 | -0.886 | 1.894828 |
| -1.53826 | -1.28973 | -0.22361 | 0 | -0.5026 | -1.09389 | -0.60445 | -0.78344 |
| 0.134575 | -0.80884 | -0.22361 | 0 | -1.45426 | -0.53784 | -0.13519 | -0.78344 |
| -1.23411 | -0.39665 | -0.22361 | 0 | -1.28123 | -0.90854 | -0.46367 | -0.78344 |
| -0.0175 | -0.46535 | -0.22361 | 0 | -0.58912 | -1.04755 | -0.60445 | -0.78344 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| -1.18342 | -0.80884 | -0.22361 | 0 | -0.93517 | -0.67685 | -0.27597 | -0.78344 |
| -0.6258 | -0.25925 | -0.22361 | 0 | -1.19472 | 1.454654 | -1.44911 | -0.40083 |
| -0.22027 | 0.015539 | -0.22361 | 0 | -1.1082 | -0.44517 | -0.36982 | -0.78344 |
| ... | ... | ... | ... | ... | ... | ... | ... |

## 3.0. Results and Discussion

In this study, we assessed the performance of two classification algorithms—Logistic Regression (LR) and K-Nearest Neighbors (KNN)—in the context of predicting subcellular protein localization. Utilizing a dataset split of 70% for training and 30% for testing, we applied Python to implement both algorithms and evaluate their effectiveness through various performance metrics, including accuracy, precision, recall, and F-score. The confusion matrices (Figures 2 and 3) illustrate how well each algorithm classified the different protein localization categories. For Logistic Regression, the model achieved an accuracy of approximately 92.86%, identifying most classes correctly, though it struggled with certain categories, reflected in lower precision and recall for some classes. In contrast, KNN outperformed with an accuracy of about 97.62%, demonstrating strong performance across all classes, as indicated by perfect precision and recall for several categories. Logistic Regression exhibited a precision of 97.5% for the most common class, but lower values for some less frequent categories, which suggests a tendency towards misclassifying certain classes.

KNN showed a marked improvement, achieving 100% precision and recall in multiple categories, underscoring its robustness in handling the dataset. The performance scores (Table 5) highlight KNN's superiority in overall effectiveness, especially in recall and F-score, indicating that KNN maintained a lower rate of false negatives.

This study contributes to the existing body of literature on protein localization classification by providing comparative insights into the performance of LR and KNN within this specific domain. While previous works have often favored more complex models, our findings suggest that KNN, with its simple yet effective approach, can outperform traditional algorithms like Logistic Regression, especially in scenarios with class imbalance. The presence of class imbalance in the dataset can affect the algorithms' performance. More advanced techniques, such as SMOTE (Synthetic Minority Over-sampling Technique), could be utilized to mitigate this issue. The comparative performance of both algorithms is captured in Figures 4 to 9 below while the performance of classification methods combining all association measures are represented in Table 5.

```
array([[39,  1,  0,  0,  0,  0,  0,  0],
       [ 1, 14,  0,  0,  0,  0,  0,  0],nm      550

       [ 0,  1,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  1,  0,  0,  0],
       [ 0,  0,  0,  0,  7,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  5,  0,  2],
       [ 0,  0,  0,  0,  0,  0,  2,  0],
       [ 0,  0,  0,  0,  0,  0,  0, 11]], dtype=int64)
```

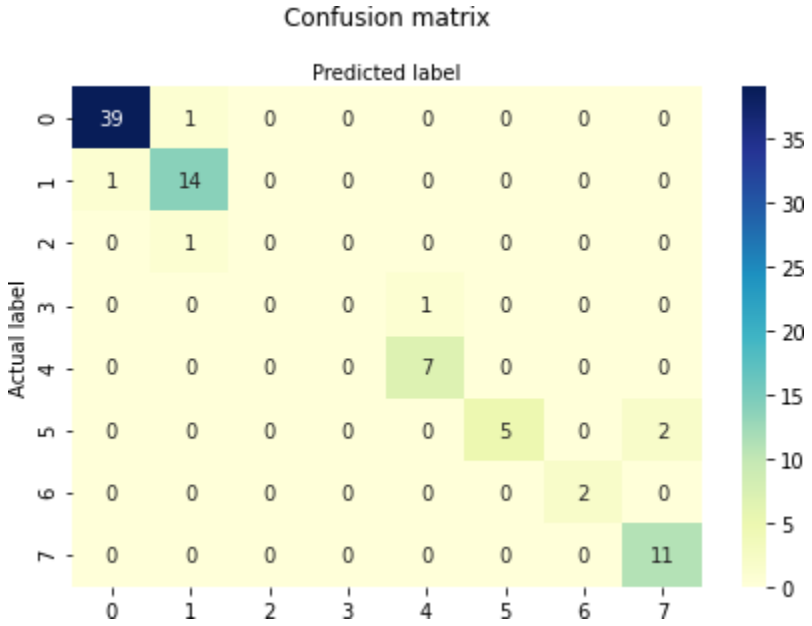


Figure 2: Confusion matrix on Logistic Regression classification

```
precision: [0.975      0.875      0.       0. 0.875      1.
 1.         0.84615385]
recall: [0.975      0.93333333 0.         0.         1.         0.71428571
 1.         1.        ]
fscore: [0.975      0.90322581 0.         0.         0.93333333 0.83333333
 1.         0.91666667]
support: [40 15  1  1  7  7  2 11]
Accuracy: 0.9285714285714286
```

```
array([[40,  0,  0,  0,  0,  0,  0,  0],
       [ 0, 15,  0,  0,  0,  0,  0,  0],
       [ 0,  1,  0,  0,  0,  0,  0,  0],
       [ 0,  0,  0,  0,  1,  0,  0,  0],
       [ 0,  0,  0,  0,  7,  0,  0,  0],
       [ 0,  0,  0,  0,  0,  7,  0,  0],
       [ 0,  0,  0,  0,  0,  0,  2,  0],
       [ 0,  0,  0,  0,  0,  0,  0, 11]], dtype=int64)
```

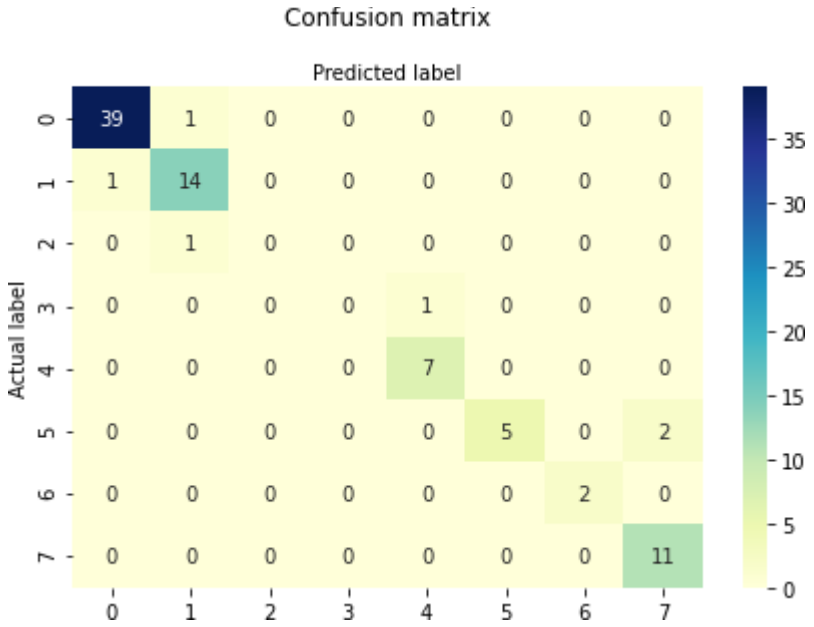


Figure 3: Confusion matrix on k neighbors classification

```
precision: [1.     0.9375 0.     0.     0.875  1.     1.     1.    ]
recall: [1. 1. 0. 0. 1. 1. 1. 1.]
fscore: [1.         0.96774194 0.         0.         0.93333333 1.
 1.         1.        ]
support: [40 15  1  1  7  7  2 11]
Accuracy: 0.9761904761904762
```

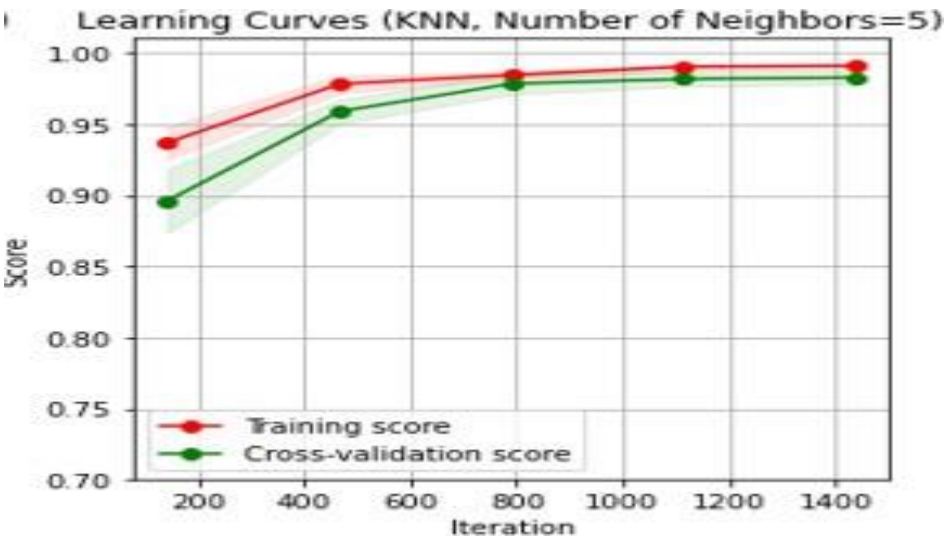


Figure 4: learning curves for KNN

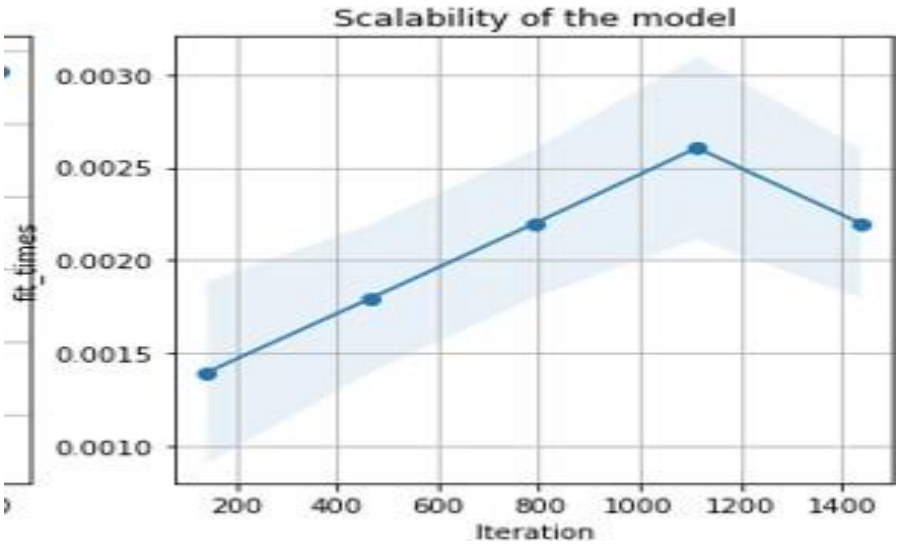


Figure 5: Scalability of the model (KNN)


Figure 6: performance  model (KNN)

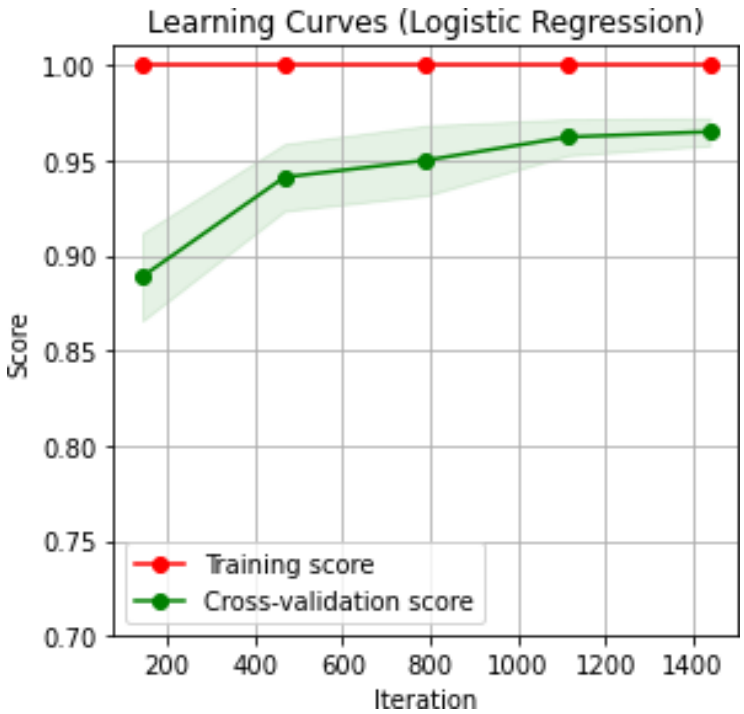


Figure 7: Learning curves for logistic regression

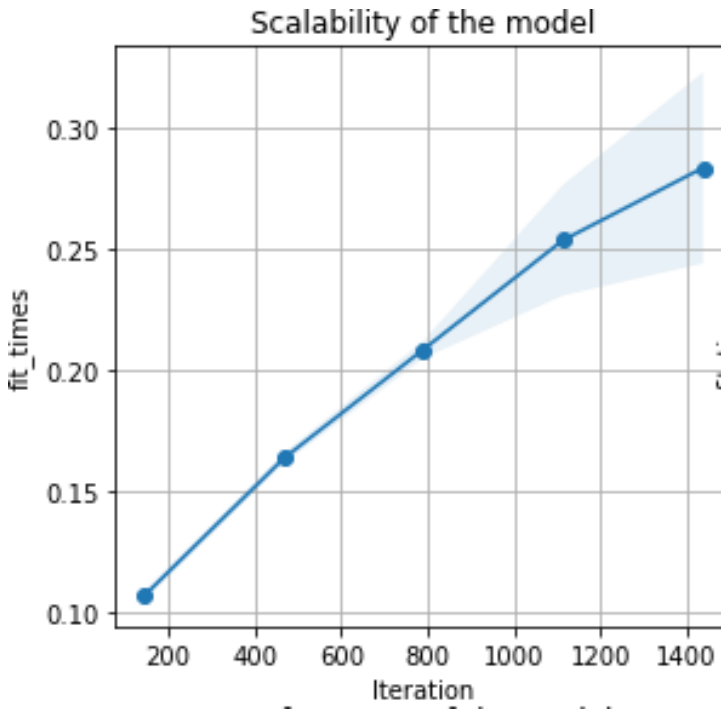


Figure 8: Scalability of Logistic regression

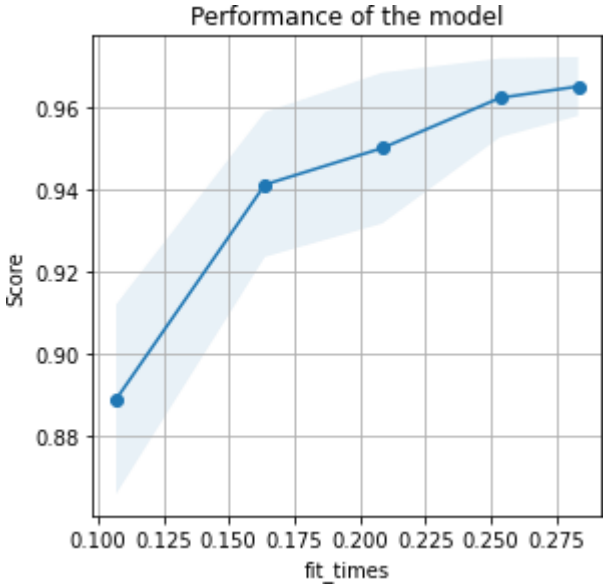

Figure 9: performance of Logistic regression model

**Table 5. Performance of Classification Methods Combining All Association Measures**

| Classifier | Precision % | Recall % | F.score% | Accuracy % |
|---|---|---|---|---|
| LR | 96% | 96% | 96% | 92% |
| KNN | 100% | 100% | 100% | 98% |

## 4.0. Conclusion

Proteins play a crucial role in various biological processes within the body. They are responsible for catalysing metabolic reactions, transporting molecules from one area of the body to another, mediating cell repair, and also form a part of our immune system. In order to elucidate a protein's function, one key piece of information is the location of the protein within the cell. Proteins perform many important tasks in living organisms, such as catalysis of biochemical reactions, transport of nutrients, and recognition and transmission of signals. The plethora of aspects of the role of any particular protein is referred to as its "function." One aspect of protein function that has been the target of intensive research by computational biologists is its subcellular localization. Proteins must be localized in the same subcellular compartment to cooperate toward a common physiological function. Predicting the location of a protein within the cell can help in elucidating its function and deducing its involvement in certain biochemical pathways. Hence this project employed, our study provides a comparative analysis of Logistic Regression (LR) and K-Nearest Neighbors (KNN) for predicting subcellular protein localization. The results demonstrate that KNN outperforms LR, achieving an accuracy of 97.62% compared to LR's 92.86%. While LR displayed high precision for the most common class, it struggled with less frequent categories, indicating a tendency for misclassification. In contrast, KNN consistently delivered 100%

precision and recall in several categories, showcasing its effectiveness in handling the dataset and minimizing false negatives.

These findings highlight the potential of KNN as a robust classifier in protein localization tasks, especially in the presence of class imbalance. The study suggests that while more complex models are often favored, simpler algorithms like KNN can provide superior performance in specific contexts. Future research may explore advanced techniques, such as SMOTE, to further enhance classification accuracy in imbalanced datasets. Overall, this work contributes valuable insights to the field of protein localization classification, encouraging the consideration of KNN as a viable alternative to traditional models.

## Reference

[1] Mazzarello P (1999). A unifying concept: the history of cell theory. Nature Cell Biology, 1:13 15.
[2] Popgeorgiev, N.; Jabbour, L.; Gillet, G (2018). “Subcellular localization and dynamics of the Bcl-2 family of proteins. Front. Cell Dev. Biol. 6, 13.
[3] Jadot M, Boonen M, Thirion J, Wang N , Xing J , Zhao C, Tannous A , Qian M, Zheng H, Everett J.K, Moore D. F, Sleat D. E, Lobel P (2017). “Accounting for Protein Subcellular Localization: A Compartmental Map of the Rat Liver Proteome”. Mol Cell Proteomics; Proteomics: Methods and Protocols, Method Molecular Biology, 1156:157{174.
[4] Wan S, Duan Y, Zou Q (2017). HPSLPred: An ensemble multi-label classifier for human protein subcellular location prediction with imbalanced source. Proteomics 2017; Proteome. Mol Cell Proteomics ; 16(2): 194-212.
[5] Cheng, X.; Xiao, X.; Chou, K-C (2017). pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics, 110(4), 231-239.
[6] Camp, R.L.; Chung, G.G.; Rimm, D.L (2002). “Automated subcellular localization and quantification of protein expression in tissue microarrays. Nat. Med., 8(11), 1323-1327.
[7] Kuo-Chen, C (2019). Artificial intelligence (AI) tools constructed via the 5-steps rule for predicting post-translational modifications. Trends Artifi. Intell. 3(1), 60-74.
[8] Horton P, Mukai Y, and Nakai K (2004). The practical bioinformati-cian. World Scientic, 1 edition, 2004.
[9] Bouziane H, Messabih B, Chouarfia A (2012). Meta-Learning for Escherichia Coli Bacteria Patterns Classification, International Conference on Web and Information Technologies (ICWIT),Proceedings of the 4th International Conference on Web and Information Technologies Sidi Bel Abbes, Algeria, April 29-30, 139-150.
[10] Jiancheng Z, Jianxin W, Wei Peng, Zhen Zhang (2015) “A Feature Selection Method for Prediction Essential Protein. Tsinghua Science & Technology 20(5):491-499
[11] Liqi L, Sanjiu Y, and Weidong X et.all (2014). Prediction of bacte-rial protein subcellular localization by incorporating variousfeatures into chou’s pseaac and a backward feature selectionapproach. Biochimie, 20(5):100–107.
[12] Muhammad I. A. J, Mohammed N. A. S, Bilal Ezz El-Din A, Hossam A. N (2021). “Predicting Protein Localization Sites in Cells using ANN”, Zaqout International Journal of Academic Information Systems Research (IJAISR) ISSN: 2643-9026 Vol. 5 Issue 3, Pages: 33-42
[13] Hoglund, A.; Donnes, P.; Blum, T.; Adolph, H.W.; Kohlbacher, O (2006). MultiLoc: prediction of protein subcellular localization using Nterminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics, 22(10), 1158-1165.