

# Predicting Protein Subcellular Localization of E. coli Bacteria Using Machine Learning Classifiers

Mamata Das<sup>1</sup>

<sup>1</sup> National Institute of Technology Tiruchirappalli, Tamil Nadu, Trichy, 620015, India

## Abstract

Protein subcellular localization refers to the specific compartments within a cell where proteins are situated, a critical aspect influencing their functions. Understanding subcellular localization is paramount in deciphering cellular processes, as proteins operate optimally within distinct cellular niches. This knowledge holds significance in areas such as cytobiology, proteomics, and drug design, as it unveils crucial insights into the intricate organization and functioning of cells. This work uses a large dataset that includes features like mcg, gvh, lip, chg, aac, alm1, alm2, and site to predict the subcellular localization of E. coli bacteria using machine learning classifiers. Classification and Regression Trees, Naive Bayes, K-Nearest Neighbors, Linear Discriminant Analysis, Logistic Regression, Support Vector Machine, Linear Support Vector Machine, Extra Trees Classifier, and Random Forest Classifier are among the classifiers that are being examined. Performance measures including recall, accuracy, precision, and F1-score are carefully assessed to give a detailed picture of each classifier's effectiveness. With an accuracy of 87.16%, precision of 85.70%, recall of 86.86%, and an F1-score of 85.77%, SVM stands out as the most effective classifier. This study adds significant knowledge to the field of microbial biology by demonstrating how machine learning may be used to forecast the subcellular location of E. coli bacteria, which has implications for more general predictive modeling applications

## Keywords

Protein, protein subcellular localization, E.coli, machine learning, microbial biology

## 1. Introduction

Eukaryotic cells exhibit intricate compartmentalization within distinct membrane-bound structures, encompassing components such as the extracellular space, plasma membrane, cytoplasm, nucleus, mitochondria, Golgi apparatus, endoplasmic reticulum (ER), peroxisome, vacuoles, cytoskeleton, nucleoplasm, nucleolus, nuclear matrix, and ribosomes. In a similar vein, bacterial cells showcase subcellular localizations discernible through cell fractionation. Essential localizations include the cytoplasm, cytoplasmic membrane (referred to as the inner membrane in Gram-negative bacteria), cell wall (typically thicker in Gram-positive bacteria), and extracellular environment. While the cytoplasm, cytoplasmic membrane, and cell wall constitute subcellular localizations, the extracellular environment stands apart. Gram-negative bacteria additionally feature an outer membrane and periplasmic space. Unlike eukaryotes, bacteria typically lack membrane-bound organelles, although exceptions like magnetosomes exist [1].

The localization of proteins within a cell is intricately tied to their functions. Proteins operate effectively only when positioned in specific subcellular compartments, underscoring the significance of studying protein localization in cytobiology, proteomics, and drug design. The prediction of protein subcellular localization through machine learning has emerged as a timely and highly engaging area within bioinformatics. This paper conducts a comprehensive review of the current research landscape surrounding protein subcellular localization prediction, focusing on four key facets. First and foremost, our initial undertaking involved the careful selection of a benchmark dataset for our study on protein subcellular localization prediction. Subsequently, we meticulously analyzed the chosen dataset, delving into its characteristics and intricacies to ensure

---

COLINS-2024: 8th International Conference on Computational Linguistics and Intelligent Systems, April 12–13, 2024, Lviv, Ukraine

✉ [dasmamata.india@mail.com](mailto:dasmamata.india@mail.com) (M. Das)

ORCID [0000-0002-5106-5571](https://orcid.org/0000-0002-5106-5571) (M. Das)

© 2024 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

a comprehensive understanding of the underlying biological information. Following this, we proceeded to select state-of-the-art machine learning models tailored to the specific task of predicting protein subcellular localization. Our model selection process considered the nuanced features of the dataset and the diverse methodologies employed by various classifiers. The culmination of our work involved a thorough analysis and comparison of the results obtained from the chosen machine learning algorithms, providing valuable insights into their performance and efficacy for the given biological prediction task.

Microbial biology, at the intersection of microbiology and computational sciences, has witnessed significant advancements with the advent of machine learning techniques. The accurate prediction of subcellular localization in bacteria, such as *Escherichia coli* (*E. coli*) [2, 3], is a critical aspect of understanding their cellular functions, metabolic pathways, and potential roles in both health and disease. *E. coli*, a well-studied bacterium, serves as an ideal model organism for such investigations due to its ubiquity in scientific research and its importance in various fields, including biotechnology and medicine.

Subcellular localization [4], referring to the specific cellular compartments or structures where proteins and biomolecules are localized, is a key determinant of their functions. Predicting the subcellular localization of *E. coli* proteins can unravel insights into its pathogenicity, virulence factors, and contribute to our understanding of its adaptation strategies in different environments.

Machine learning, a subset of artificial intelligence, has proven to be an invaluable tool in deciphering complex biological data [5]. In this study, we harness the power of diverse machine learning classifiers to predict the subcellular localization of *E. coli* bacteria. The dataset employed encompasses a range of biological features, including *mcg*, *gvh*, *lip*, *chg*, *aac*, *alm1*, *alm2*, and *site*, each serving as a molecular signature influencing subcellular localization.

The classifiers selected for evaluation comprise a comprehensive set, including Logistic Regression (LR) [6], Linear Discriminant Analysis (LDA) [7], K-Nearest Neighbors (KNN) [8], Classification and Regression Trees (CART) [9], Naive Bayes (NB) [10], Support Vector Machine (SVM) [11], Linear Support Vector Machine (L-SVM) [12], Extra Trees Classifier (ETC) [13], and Random Forest Classifier (RFC) [14]. These classifiers are chosen for their diverse methodologies and suitability for different types of data.

The evaluation of performance metrics, such as accuracy, precision, recall, and F1-score, is central to our analysis. These metrics provide a holistic view of each classifier's ability to make accurate predictions, balance precision and recall, and effectively discern subcellular localization patterns in *E. coli*.

This study not only contributes to the growing body of knowledge in microbial biology but also underscores the potential of machine learning in unraveling the complexities of bacterial subcellular organization. The outcomes hold implications for advancements in predictive modeling, offering a nuanced understanding of *E. coli* biology and paving the way for broader applications in microbial research and biotechnology.

## 2. Related Works

The exploration of subcellular localization prediction in microorganisms, particularly bacteria like *Escherichia coli* (*E. coli*), has been a focal point in bioinformatics and computational biology. The task involves predicting the cellular compartments or locations within a cell where proteins are likely to reside. Several studies have delved into this domain, employing diverse methodologies ranging from traditional bioinformatics approaches to more contemporary machine learning techniques. In this section, we will review and analyze the most significant contributions in the field of subcellular localization prediction over the past decade, focusing specifically on work published in the last 10 years.

Developing on the previous Hum-mPLOC predictor, the enhanced Hum-mPLOC 2.0 tackles challenges in predicting subcellular localization of human proteins [17], especially those with multiplex characteristics. Unlike its predecessor, Hum-mPLOC 2.0 eliminates the need for protein

accession numbers, making it applicable to proteins without such identifiers. Additionally, it incorporates functional domain and sequential evolution information through an ensemble classifier, resulting in a substantially improved prediction capability. The freely accessible web server for Hum-mPloc 2.0 offers an efficient solution to address these shortcomings in subcellular localization prediction.

CELLO2GO [18] is a web-based system offering a comprehensive screening of targeted proteins, providing gene ontology (GO)-type categories, and subcellular localization information. The platform utilizes BLAST homology searching and CELLO localization prediction, combining these approaches to generate detailed GO annotations and predict subcellular localization based on the identified homologous sequences. CELLO2GO's output includes informative pie charts summarizing the functional annotations, making it a valuable tool for complex subcellular system research by integrating CELLO and BLAST functionalities into a user-friendly platform.

The author has proposed a novel SVM-based approach, MultiLoc [20], with the aim of enhancing proteomic functional annotation. This method integrates N-terminal targeting sequences, amino acid composition, and protein sequence motifs for comprehensive subcellular localization prediction. Through comparisons with existing methods, the study demonstrates improved predictions based on N-terminal targeting sequences using our method, TargetLoc. MultiLoc exhibits superior or comparable performance to specialized methods focused on fewer localizations or specific organisms when predicting major eukaryotic subcellular localizations.

The paper [15] explores the spatial organization of proteins in bacterial cells, highlighting specific locations where proteins congregate. It emphasizes the role of cellular shapes, self-assembly, and designated sites in guiding proteins to their functional positions. Using examples such as FtsZ for cell division and proteins involved in chemotaxis and spore formation, the paper elucidates how proteins contribute to vital processes, including growth, cell cycle regulation, and behavioral changes in bacterial cells. The authors anticipate advancements in microscopy and tracking techniques to unveil intricate details of protein movement and function in bacteria, underscoring the significance of understanding protein localization and suggesting avenues for further research [25, 44].

The paper [28] introduces LOCALIZER, a novel computational method designed to predict plant and effector protein localization accurately within chloroplasts, mitochondria, and nuclei. Exhibiting enhanced accuracy compared to existing methods, LOCALIZER proves invaluable for prioritizing effector candidates and sheds light on subcellular localization dynamics in plant-pathogen interactions.

The paper [43] introduces COMPARTMENTS, a comprehensive tool serving as a knowledge hub for protein subcellular localization. By aggregating data from diverse sources and utilizing text mining, it continuously updates with confidence scores, simplifying information visualization through cell diagrams, categorizing evidence, and assigning reliability scores, aiming to facilitate researchers in comprehending and comparing protein location information within cells.

The paper [23, 30] presents a support vector machine method for precise protein subcellular localization prediction using amino acid sequences. This method maintains effectiveness despite errors in the initial protein sequence, and comparative analysis highlights its superiority over other methods, proving valuable for large-scale genetic information analysis and contributing significantly to biology and genetics research.

The paper introduces Dynamic Organellar Maps [42], enabling comprehensive mapping of protein translocation in HeLa cells with over 8700 proteins, providing detailed spatial and abundance information for quantitative analysis of cell anatomy and organellar composition, both statically and dynamically in response to stimuli like EGF. This method allows proteome-wide exploration of physiological protein movements without requiring process-specific reagents, offering broad applicability in cell biology.

The author has proposed a stacked ensemble-based deep learning model [41] for the multi-label classification of protein subcellular localization, showcasing superior performance compared to existing approaches in the Human Protein Atlas database.

### 3. Proposed Methodology

#### 3.1. Dataset

We have used the E.coli (*Escherichia coli*) bacteria dataset, taken from UC Irvine machine learning repository database. The data is containing 336 instances and 7 features. Table 1 displays the attribute information of the dataset, with the first column containing attribute names and the second column providing descriptions. The distribution of localization sites is presented in Table 2, where the first column denotes location names and the second column indicates the count for each specific location.

**Table 1**  
**Attribute information [31, 32]**

| Sequence Name | Accession number for the SWISS-PROT database   |
|---------------|--|
| mcg           | McGeoch's method for signal sequence recognition.  |
| gvh           | von Heijne's method for signal sequence recognition  |
| lip           | von Heijne's Signal Peptidase II consensus sequence score. Binary attribute.                         |
| chg           | Presence of charge on N-terminus of predicted lipoproteins. Binary attribute.                        |
| aac           | score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins. |
| alm1          | score of the ALOM membrane spanning region prediction program.                                       |
| alm2          | score of ALOM program after excluding putative cleavable signal regions from the sequence.           |

**Table 2**  
**Distribution of localization site**

| Locations   | Count |
|---|-------|
| cytoplasm (cp)                                    | 143   |
| inner membrane without signal sequence (im)       | 77    |
| periplasm (pp)                                    | 52    |
| inner membrane, uncleavable signal sequence (imU) | 35    |
| outer membrane (om)                               | 20    |
| outer membrane lipoprotein (omL)                  | 5     |
| inner membrane lipoprotein (imL)                  | 2     |

The statistical description of the Ecoli bacteria data reveals important insights about the dataset in Table 3. There are 336 observations for each attribute, indicating a consistent dataset size. The mean values provide an average measure for each attribute. Notably, the means for the attribute's mcg, gvh, lip, chg, aac, alm1, alm2, and site vary. The standard deviation provides a measure of the dispersion or spread of the data. A lower standard deviation suggests that the data points tend to be closer to the mean. The minimum and maximum values highlight the range of each attribute. For example, the site attribute has a minimum value of 1.000 and a maximum value of 8.000, indicating the range of classes or categories.

**Table 3**  
**Statistical description of E.coli bacteria data**

|       | mcg    | gvh    | lip    | chg    | aac    | alm1   | alm2   | site   |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|
| count | 335.00 | 335.00 | 335.00 | 335.00 | 335.00 | 335.00 | 335.00 | 335.00 |
| mean  | 0.500  | 0.5017 | 0.496  | 0.501  | 0.500  | 0.501  | 0.500  | 2.245  |
| std   | 0.195  | 0.148  | 0.089  | 0.027  | 0.123  | 0.216  | 0.210  | 1.443  |
| min   | 0.000  | 0.160  | 0.480  | 0.500  | 0.000  | 0.030  | 0.000  | 1.000  |

|     |       |       |       |       |       |       |       |       |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|
| max | 0.890 | 1.000 | 1.000 | 1.000 | 0.880 | 1.000 | 0.990 | 8.000 |
|-----|-------|-------|-------|-------|-------|-------|-------|-------|

Note: The Table 3 provides a statistical summary of various attributes (mcg, gvh, lip, chg, aac, alm1, alm2, site) in the Ecoli bacteria dataset, including the count, mean, standard deviation (Std), minimum (Min), and maximum (Max) values for each attribute.

### 3.2. Model

The classifiers selected for assessment are a broad and varied group, each chosen for its own approach and suitability for different kinds of data. The ensemble comprises a model of the likelihood of class membership called Logistic Regression (LR); Classification and Regression Trees (CART), which uses tree-like models to make judgments; K-Nearest Neighbors (KNN), a non-parametric technique based on similarity measurements; Linear Discriminant Analysis (LDA), which looks for the linear combinations of characteristics that best discriminate classes; Extra Trees Classifier (ETC), which uses an ensemble of decision trees with random feature splits; Support Vector Machine (SVM), which builds hyperplanes for optimal class separation; Linear Support Vector Machine (L-SVM), an SVM variant intended for linearly separable data; and as well as the Random Forest Classifier (RFC), an ensemble technique that combines forecasts from several decision trees. This broad selection guarantees a comprehensive analysis, taking into account the advantages and flexibility of each classifier to different features in the dataset being analyzed. Incorporating classifiers with disparate underlying concepts enhances the comprehensiveness of the research by providing insights into their relative performance and appropriateness for various types of data.

### 3.3. Performance

In our study, a critical insight emerges-acknowledging that not all correct or incorrect matches carry the same significance. Relying on a singular metric falls short of providing a comprehensive assessment of classification performance. Consequently, we have opted for a multi-faceted approach, utilizing accuracy, recall, precision, and F1 score as performance metrics, which will be elaborated upon in the subsequent section. Table 4 presents the comprehensive set of performance metrics utilized in our study.

#### 3.3.1. Accuracy

Accuracy stands out as the most instinctive performance measure, representing the ratio of correctly predicted observations to the total number of observations. A model is deemed optimal when achieving high accuracy or nearing perfection [33, 34]

#### 3.3.2. Precision

Put simply, precision can be conceptualized as a gauge of a classifier's precision-the extent to which identifications are accurate. It reflects the ratio of correctly predicted positive instances to the total predicted positive instances [37, 38]. A lower precision value may suggest a higher count of False Positives

**Table 4**  
**Performance metrics**

| Performance Metrics | Description                                       |
|---------------------|---|
| Accuracy            | $(TP + TN) / (TP+TN+PF+FN)$                       |
| Precision           | $TP / (TP+FP)$                                    |
| Recall              | $TP / (TP+FN)$                                    |
| F1-Score            | $(2 * Precision * Recall) / (Precision + Recall)$ |

### 3.3.3. Recall

Recall serves as a metric reflecting a classifier's comprehensiveness, revealing the proportion of actual positives correctly identified by the predictive model [35, 36, 37]. It is the ratio of correctly predicted positive instances, encompassing both true positives and false negatives. Additionally, known as Sensitivity, a low recall value indicates a notable count of False Negatives.

### 3.3.4. F1-Score

In scenarios with imbalanced class distribution, F1 emerges as a suitable performance metric. Being the weighted average of Precision and Recall, this score incorporates considerations for both false negatives and false positives [39, 40]. Alternatively, one might express that the F1 score encapsulates the equilibrium between precision and recall. An effective information retrieval or text classification classifier is anticipated to yield high or close-to-high values for precision, recall, and F1 score.

## 4. Results and Analysis

### 4.1. Execution environments

Our experimental setup utilized a Lenovo ThinkPad E14 Ultrabook operating on the Windows 10 Professional 64-bit system, equipped with a 10th Generation Intel Core i7-10510U Processor. The processor operates at a clock speed of 1.8 GHz, and the system is configured with a 16 GB DDR4 memory size.

### 4.2. Results

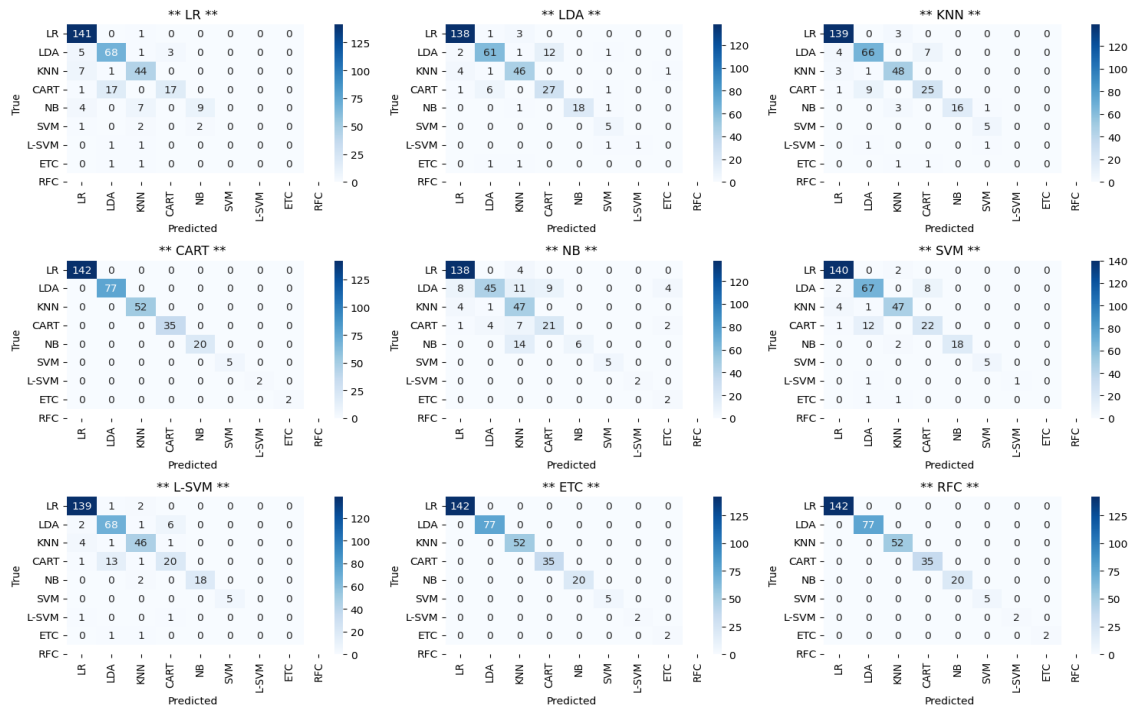
The performance classification results for the E. coli dataset reveal distinctive characteristics of various machine learning classifiers. These findings are crucial in understanding the strengths and limitations of each model. Table 4 summarizes the performance of classifiers such as LR, LDA, KNN, CART, NB, SVM, L-SVM, ETC, and RFC. Each classifier is evaluated based on Accuracy, Precision, Recall, and F1-Score. Figure 1 visually represents the confusion matrices of different machine learning models, offering insights into their performance in predicting the subcellular localization of E. coli bacteria. These matrices provide a comprehensive overview, detailing true positive, true negative, false positive, and false negative predictions, facilitating a thorough analysis of the effectiveness of each classifier in the prediction task. Table 5 presents a summary of the classification performance.

Support Vector Machine (SVM) emerges as a top performer, achieving the highest accuracy of 87.16%. This indicates the SVM's robust ability to correctly classify instances in the Ecoli dataset. The SVM also demonstrates commendable precision (85.70%) and recall (86.86%), striking a well-balanced trade-off between false positives and false negatives, as evident from the high F1-score of 85.77%.

Linear Discriminant Analysis (LDA) showcases balanced performance across multiple metrics, standing out with an accuracy of 86.84%. LDA's precision (88.08%) and recall (86.90%) contribute to a high F1-score of 86.98%, reinforcing its reliability in correctly identifying positive instances.

K-Nearest Neighbors (KNN) aligns closely with LDA, achieving an accuracy of 86.87%. While KNN's precision (84.00%) and recall (84.80%) are slightly lower than LDA, the F1-score remains robust at 83.87%, highlighting its competence in classification.

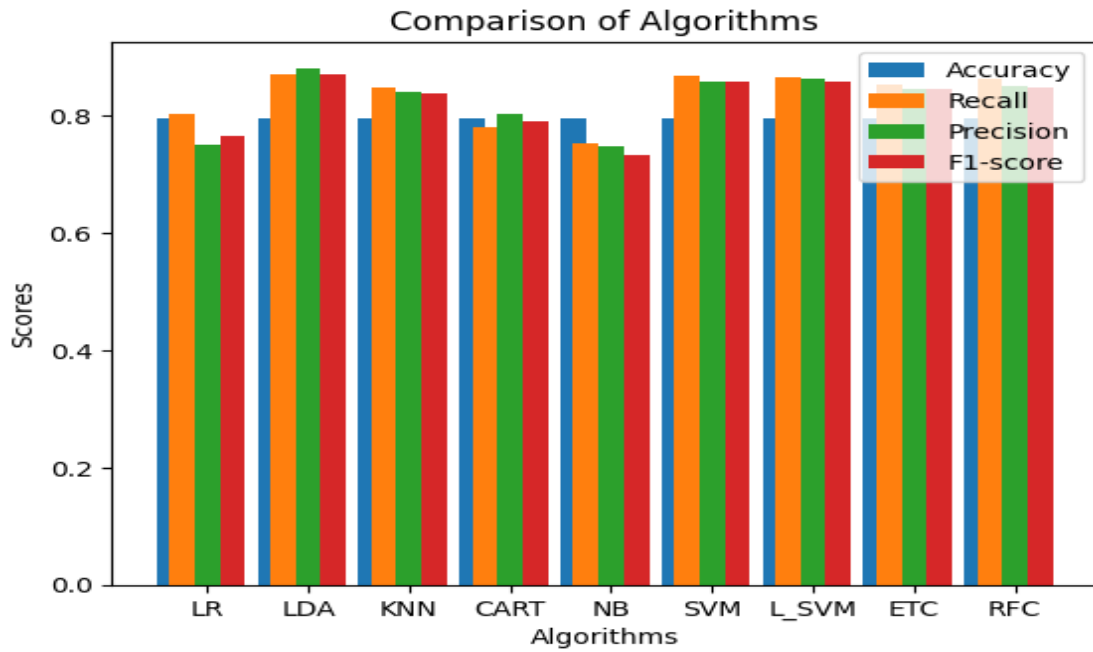
Random Forest Classifier (RFC) delivers competitive results with an accuracy of 86.25%. RFC exhibits high precision (86.32%) and recall (84.49%), striking a balance reflected in the F1-score of 84.63%.



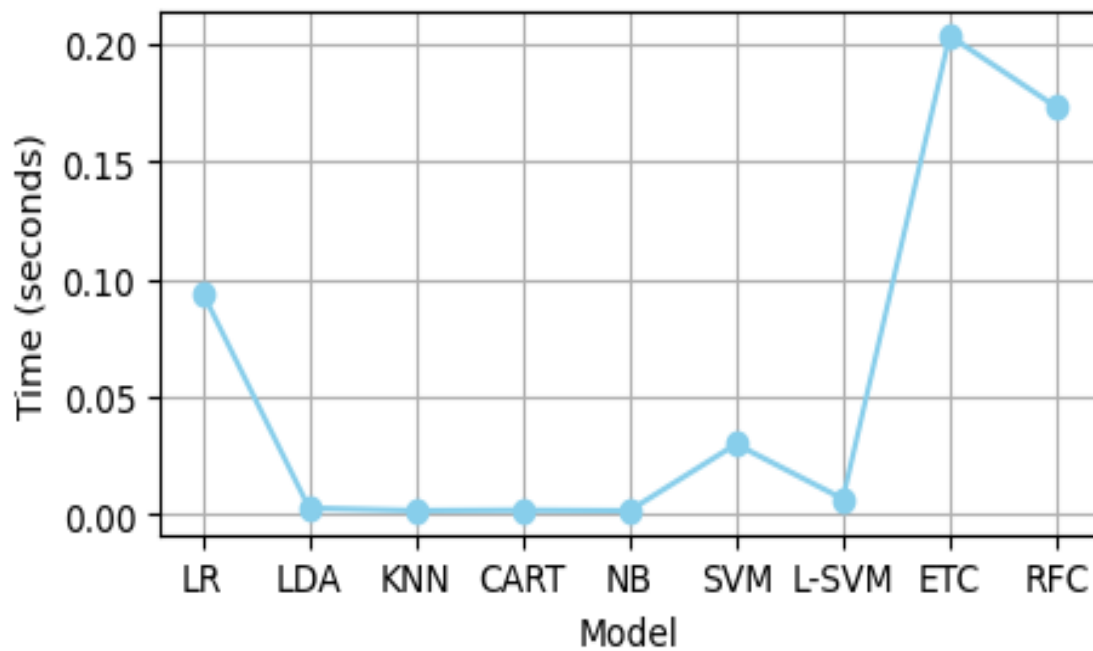
**Figure 1:** Confusion matrices for the models LR, LDA, KNN, CART, NB, SVM, L\_SVM, ETC, and RFC

**Table 5**  
**Performance classification summary for the E.coli dataset**

| Classifier | Accuracy | Precision | Recall   | F1-Score |
|------------|----------|-----------|----------|----------|
| LR         | 0.793850 | 0.749978  | 0.803000 | 0.764240 |
| LDA        | 0.868449 | 0.880774  | 0.869000 | 0.869843 |
| KNN        | 0.868717 | 0.840020  | 0.848000 | 0.838716 |
| CART       | 0.794118 | 0.808639  | 0.798000 | 0.791203 |
| NB         | 0.752317 | 0.748441  | 0.753000 | 0.733291 |
| SVM        | 0.871569 | 0.857038  | 0.868627 | 0.857711 |
| L-SVM      | 0.859626 | 0.863019  | 0.865597 | 0.856470 |
| ETC        | 0.865597 | 0.838753  | 0.845009 | 0.844410 |
| RFC        | 0.862478 | 0.863171  | 0.844920 | 0.846348 |



**Figure 2:** Algorithmic performance comparison on the E. coli dataset



**Figure 3:** Algorithmic training times comparison on the E. coli dataset

Logistic Regression (LR) and Classification and Regression Trees (CART) show comparable performance with accuracies of 79.39% and 79.41%, respectively. However, LR demonstrates a precision-recall trade-off, leading to a lower F1-score of 76.42%. CART, on the other hand, maintains a more balanced F1-score of 79.12%.

Naive Bayes (NB) exhibits a slightly lower performance with an accuracy of 75.23%. The model's precision (74.84%) and recall (75.30%) contribute to an F1-score of 73.33%, indicating potential challenges in correctly classifying positive instances.

Figure 2 illustrates a comparative analysis of algorithmic performance and training times across various machine learning models on the E. coli dataset.



## 5. Conclusion

The investigation into predicting subcellular localization of *E. coli* bacteria through machine learning classifiers yields valuable insights into the strengths and nuances of various models. Notably, Support Vector Machine (SVM) emerges as a standout performer with high accuracy, precision, recall, and F1-score, showcasing its robust predictive capabilities. Linear Discriminant Analysis (LDA) and K-Nearest Neighbors (KNN) also demonstrate commendable performances, emphasizing the importance of selecting classifiers tailored to specific application requirements. The comprehensive evaluation, considering precision, recall, and F1-score alongside accuracy, provides a holistic understanding of classifier effectiveness in real-world scenarios. Beyond the specific application to *E. coli*, these findings contribute to microbial biology, illustrating the potential of machine learning in subcellular localization prediction with implications for broader applications in microbial research and biotechnology.

In conclusion, this study advances our understanding of *E. coli* biology while providing valuable insights into the landscape of microbial research. The future trajectory of this work involves exploring advanced visualization techniques, hyperparameter tuning, and incorporating larger datasets for a more extensive evaluation of classifier generalization across diverse microbial species. Ensemble methods and deep learning approaches present promising avenues for further refinement, and the integration of biological context could enhance classifiers' interpretability. Ongoing advancements in machine learning and computational biology offer exciting opportunities for refining predictive capabilities in microbial subcellular localization [26], positioning this research at the forefront of interdisciplinary advancements.

## 6. Future Works

Anticipating the subcellular location of microorganisms, specifically *E. coli* bacterium, offers numerous avenues for investigation and advancement in the future. First, to capitalize on the advantages of several classifiers and maybe improve overall prediction performance, the integration of ensemble techniques [16], like stacking or boosting, should be researched. Investigating the use of deep learning models-like neural networks-could provide a more comprehensive comprehension of intricate correlations seen in biological data, opening the door to predictions that are more correct. Moreover, it is imperative that machine learning models prioritize interpretability and explainability, particularly when applied to biological research [46]. In the future, the models' biological relevance could be increased by using methods such as attention mechanisms or SHAP (SHapley Additive exPlanations) values to identify the features affecting predictions. Furthermore, domain-specific information like functional annotations or protein-protein interactions [45] may help produce predictions that are more context-aware. Growing the dataset to include a wider variety of microbial species and subcellular compartments would offer a more thorough assessment of classifier generalization as high-throughput technologies continue to progress [24, 27]. Furthermore, the creation of user-friendly web servers [19, 21] or applications built on the verified models may enable these predictive tools to be more widely accessible and used by the scientific community. Additionally, our study can be expanded using the Markov clustering algorithm (MCL) to analyze patterns within subcellular localization data. Employing MCL enables the identification of protein groups sharing similar localization patterns [22, 29], enhancing insights into cellular organization and protein functions. Utilizing TF-IDF [47] for predicting the cellular location of proteins represents a promising avenue for improving accuracy and reliability in subcellular localization.

## References

- [1] S. Dirk, Molecular analysis of a subcellular compartment: the magnetosome membrane in *Magnetospirillum gryphiswaldense*, *Arch Microbiol.* 181(2004):1-7.

- [2] O. Tenaillon, D. Skurnik, B. Picard, E. Denamur, The population genetics of commensal *Escherichia coli*, *Nature Reviews. Microbiology* 8.3(2010): 207–17. doi:10.1038/nrmicro2298.
- [3] P. Singleton, *Bacteria in Biology. Biotechnology and Medicine* (5th ed.). Wiley,1999
- [4] C. Yu, Y. Chen, C. Lu, J. Hwang, Prediction of protein subcellular localization, *Proteins: Structure, Function, and Bioinformatics* 64.3(2006):643—651
- [5] J. Koza and F. Bennett, D. Andre, M. Keane, Automated design of both the topology and sizing of analog electrical circuits using genetic programming, *Artificial intelligence in design'96* (1996):151—170
- [6] J. Tolles, W. Meurer, Logistic Regression Relating Patient Characteristics to Outcomes, *JAMA*. 316.5(2016): 533–4. doi:10.1001/jama.2016.7653
- [7] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Interscience. 2004
- [8] J. L. Hodges, *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*. (1951)
- [9] W. Loh, *Classification and regression trees*, *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1.1(2011):14-23
- [10] G. Webb, E. Keogh, R. Miikkulainen, Naive Bayes, *Encyclopedia of machine learning* 15(2010):713—714
- [11] A. Widodo, B. Yang, Support vector machine in machine condition monitoring and fault diagnosis, *Mechanical systems and signal processing*. 21(2007):2560—2574
- [12] S. Paul, C. Boutsidis, M. Magdon-Ismail, P. Drineas, Random projections for linear support vector machines, *ACM Transactions on Knowledge Discovery from Data (TKDD)* 8(2014): 1—25
- [13] D. Baby, S. Devaraj, J. Hemanth, thers, Leukocyte classification based on feature selection using extra trees classifier: Atransfer learning approach, *Turkish Journal of Electrical Engineering and Computer Sciences* 29(2021): 2742—2757
- [14] M. Pal, Random forest classifier for remote sensing classification, *International journal of remote sensing* 26(2005): 217—222
- [15] K. Chou, H. Shen, Large-scale plant protein subcellular location prediction, *Journal of cellular biochemistry* 100(2007): 665—678
- [16] K. Chou, H. Shen, Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization, *Biochemical and biophysical research communications* 347(2006): 150—157
- [17] H. Shen, K. Chou, A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0, *Analytical biochemistry* 394(2009): 269—274
- [18] C. Yu, C. Cheng, W. Su, K. Chang, S. Huang, J. Hwang, C. Lu, CELLO2GO: a web server for protein subCELLular LOcalization prediction with functional gene ontology annotation, *PloS one* 9(2014): e99368
- [19] C. Savojardo, P. L. Martelli, P. Fariselli, G. Profiti, R. Casadio, BUSCA: an integrative web server to predict subcellular localization of proteins, *Nucleic acids research* 46(2018): W459--W466
- [20] A. Hoglund, P. Donnes, T. Blum, H. Adolph, O. Kohlbacher, MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition, *Bioinformatics* 22(2006): 1158-1165
- [21] K. Chou, H. Shen, Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms, *Nature protocols* 3.2(2008):153-162
- [22] M. Das, P. Alphonse, K. Selvakumar, Markov clustering algorithms and their application in analysis of PPI network of malaria genes, *IDAACS* 2(2021):855—860
- [23] A. Garg, M. Bhasin, G. Raghava, Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search, *Journal of biological Chemistry* 280(2005):14427—14432
- [24] T. Goldberg, M. Hecht, T. Hamp, T. Karl, G. Yachdav, N. Ahmed, A. Nadeem, U. Altermann, others, LocTree3 prediction of localization, *Nucleic acids research* 42(2014): W350--W355

- [25] M. Bhasin, A. Garg, G. Raghava, PSLpred: prediction of subcellular localization of bacterial proteins, *Bioinformatics* 21(2005):2522—2524
- [26] H. Shen, K. Chou, Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites, *Biochemical and biophysical research communications* 355.4(2007): 1006—1011
- [27] R. Nair, B. Rost, Mimicking cellular sorting improves prediction of subcellular localization, *Journal of molecular biology* 348(2005):85—100
- [28] J. Sperschneider, A. Catanzariti, K. DeBoer, B. Petre, D. Gardiner, others, LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell, *Scientific reports* 7(2017):44598
- [29] M. Das, P. Alphonse, K. Selvakumar, An analytical study of COVID-19 dataset using graph-based clustering algorithms, *Smart Intelligent Computing and Applications, SCI-2021* 1(2022): 1—15
- [30] K. Chou, Y. Cai, Using functional domain composition and support vector machines for prediction of protein subcellular location, *Journal of biological chemistry* 277(2002):45765—45769
- [31] N. Kenta, K. Minoru, Expert Sytem for Predicting Protein Localization Sites in Gram-Negative Bacteria, *PROTEINS: Structure, Function, and Genetics* 11(1991):95-110
- [32] N. Kenta, K. Minoru, A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells, *Genomics* 14(1992): 897-911
- [33] BSB ISO, Accuracy (trueness and precision) of measurement, *International standard ISO 5725(1998)*: 1994
- [34] J. Swets, Measuring the accuracy of diagnostic systems, *Science* 240(1988): 1285—1293
- [35] D. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, *arXiv preprint arXiv:2010.16061* 2 (2020): 37–63
- [36] S. Campana, Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods, *Journal of fish biology* 59(2001): 197—242
- [37] M. Buckland, F. Gey, The relationship between recall and precision, *Journal of the American society for information science* 45(1994): 12—19
- [38] P. Franti, R. Mariescu-Istodor, Soft precision and recall, *Pattern Recognition Letters* 167(2023): 115--121
- [39] Y. Sasaki, others, The truth of the F-measure, *Teach tutor mater* 1(2007):1—5
- [40] D. Chicco, G. Jurman, The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation, *BMC genomics* 21(2020):1—13
- [41] S. Aggarwal, S. Gupta, D. Gupta, Y. Gulzar, others, An artificial intelligence-based stacked ensemble approach for prediction of protein subcellular localization in confocal microscopy images, *Sustainability* 15.2(2023):1695
- [42] D. Itzhak, S. Tyanova, J. Cox, G. Borner, Global, quantitative and dynamic mapping of protein subcellular localization, *elife* 5(2016):e16950
- [43] J. X. Binder, S. Pletscher-Frankild, S. Kalliopi, S. O'Donoghue, R. Schneider, others, COMPARTMENTS: unification and visualization of protein subcellular localization evidence, *Database* 2014(2014):bau012
- [44] D. Rudner, R. Losick, Protein subcellular localization in bacteria, *Cold Spring Harbor perspectives in biology* 2.4(2010):a000307
- [45] M. Das, K. Selvakumar, P. Alphonse, Analyzing and Comparing Omicron Lineage Variants Protein--Protein Interaction Network Using Centrality Measure, *SN Computer Science* 4.3(2023):299
- [46] W. Lin, J. Fang, X. Xiao, K. Chou, iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins, *Molecular BioSystems* 9(2013):634-644
- [47] M. Das, P. Alphonse, K. Selvakumar, A comparative study on tf-idf feature weighting method and its analysis using unstructured dataset, *COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems* (2021)