Predicting Cellular Protein localization Sites on Ecoli's Minimal Dataset using a Comparison of Machine Learning Techniques

Safari Yonasi
Mbarara University of Science and Technology
Department of IT

Rose Nakasi
Busitema University
Department of Computer Science

Yashik Singh
University Of KwaZulu-Natal
Department of Tele-Health

ABSTRACT

Several Machine Learning Classification Techniques have been applied in predicting Protein Localization sites of E.coli using a number of techniques. However, research done is limited to no prediction of Localization sites of Proteins on Ecoli′s minimal dataset with the most informative features obtained using different feature selection techniques. This study investigated several Machine learning Classification and Feature Selection Techniques as applied on Ecoli′s minimal dataset. The implementation of classifiers aided in predicting localization sites of E.coli′s minimal subset using its informative features obtained by feature selection techniques. Results were achieved in four parts including; (Data Collection, Cleaning and Preprocessing), Feature selection where the most informative features are selected, Classification where prediction of the localization of proteins is done and then Evaluation of the Classifiers to assess their performance using a number of measures including Accuracy from Cross-validation, and AUROCC to enable in recommending the best Classifier at the end. Among the Classifiers used, Extra Tree Classifier and Gradient Boosting are seen to be the best at performance followed by Random forest as seen from Precision, Recall and F-measure scores. AdaBoost is the worst at 83%.

General Terms

Machine learning Classification Techniques

Keywords

Predicting, Ensemble and Non Ensemble Classifiers, and Machine Learning Techniques

1. INTRODUCTION

E.coli (Escherichia coli) is a type of coliform bacteria commonly found in the intestines of humans and warm-blooded animals. The organism is most frequently responsible for urinary tract infections and it is the bacterium most often implicated in the cause of diarrhea in people traveling overseas [10]. The prevalence of multi drug-resistant E. coli strains is increasing worldwide principally due to the spread of mobile genetic elements, such as plasmids [1]. According to a Study carried out by Sabir [29], several antibiotics including penicillin, amoxicillin, cefotaxime, beta lactam antibiotics including cefotaxime, ceftazidime, cephradin, cefuroxime, cephradin, Ceftrioxone among others are used when treating Ecoli infections. However, some of the E.coli species are resistant to some of these antibiotics such as penicillin and amoxicillin. This indicates a cautious use of these antibiotics for the treatment of urinary tract infections [29].

More than a third of the world's population has no access to essential drugs with more than half of this group of people living in the poorest regions of Africa and Asia[30].
Drugs offer a simple, cost-effective solution to many health problems, provided they are available and affordable. The objective of medicine is to address people's unavoidable needs for emotional and physical healing [4][7]. The discipline has evolved over millennial by exploiting natural products in their environments, and more recently by developing and validating therapeutic and preventive approaches using the scientific method.

Public health and medical practices have now advanced to a point at which people can anticipate and even feel entitled to lives that are longer and of better quality than ever before in human history. Yet despite the pervasiveness, power, and promise of contemporary medical science, large segments of humanity either cannot access its benefits or choose not to do so. More than 80 percent of people in developing nations can barely afford the most basic medical procedures, drugs, and vaccines [2].

Elucidating the protein function is very relevant for genome annotation and search for novel vaccine or drug discovery. The most reliable way to determine protein structure or function is by direct experimentation [5]. Unfortunately, it is laborious, expensive and time consuming to use purely experimental techniques though it is accurate [8]. However, the amino acid sequence of a protein usually provides crucial indication to its cellular localization sites. This has been used in areas such as Bioinformatics in predicting which part in a cell a given protein is transported to, where an amino acid sequence of the protein is given as an input [34][31].
Proteins are transported to various localization sites within a cell in order to function properly. The cellular localization site of a protein affects its potential functionality as well as its accessibility to drug treatments. Fortunately, the information needed for correct localization is generally found in the protein sequence itself. On the other hand, sequenced genomic data is experiencing an exponential

increase in recent years due to maturation of High-Throughput sequencing techniques. Thus, many computational methods have been proposed to try to set up the link between a protein sequence and its cellular location. These include McGeoch's method for signal sequence recognition, discriminant analysis of the amino acid content of outer membrane and periplasmic proteins, among others. However, each of these methods can only deal with one protein category by giving the probability of a sequence being a membrane protein, or deciding whether it is a nucleus protein or not. Therefore, for a new protein sequence on which people have no pre-knowledge, the only way to decide its localization site is to check all available methods to get a sense. However, people still need to judge between these results to decide which method is more reliable, what is the cutoff probability for it to be safe to say a protein is in a certain cellular localization site but not in other sites. Thus, it is in a great need to develop a comprehensive system, integrating protein sequence-derived data and prediction results from all the methods described above [34].

Scientists involved in the area of proteomics are currently seeking integrated, customized and validated research solutions to better expedite their work in proteomics analyses and drug discoveries. Through the development of new approaches in computer science, coupled with an increased dataset of proteins of known localization, computational tools can now provide fast and accurate localization predictions for many organisms [27]. It has been showed that a variety of machine learning methods can be used for this purpose and they seem to be a more realizable and very promising solution [8][34][5].

The first approach for predicting the localization sites of proteins from their amino acid sequences was a rule based expert system PSORT, and then the use of a probabilistic model which could learn its parameters from a set of training data, improved significantly the classification accuracy. Later, the use of standard classifiers achieved higher accuracy. Among these algorithms, k-Nearest Neighbors (k-NN), binary decision tree and nave Bayesian classifier [5].

Several other scholars have since attempted to investigate this effect using classifiers such as feed-forward neural networks and ensembles [27][3].

This piece of research aims at developing an approach that analyzes different classification, and feature selection techniques and suggests the best methods in Predicting Cellular Protein localization Sites on Ecoli's minimal dataset. This minimal dataset is a representative of Ecoli's most informative features.

2. RELATED WORK IN PREDICTING PROTEIN LOCALIZATION SITES IN E.COLI

Predicting Protein localization Sites in E.coli has been attempted before by other researchers.

P Horton et.al [14] 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 neighbors 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.

Others [12] have 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 [12]. 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. These minority classes are the most difficult to classify.

Scientists involved in the area of Proteomics are currently seeking integrated, customized and validated research solutions to better expedite their work in Proteomics analyses and drug discoveries. Some drugs and most of their cell targets are proteins, because proteins dictate biological phenotype. In this context, the automated analysis of protein localization is more complex than the automated analysis of DNA sequences; nevertheless the benefits to be derived are of same or greater importance. In order to accomplish this target, the right choice of the kind of the methods for these applications, especially when the data set is drastically imbalanced, is very important and crucial. Performance of some commonly used classifiers is investigated, for example the K nearest neighbors and feed-forward neural networks with and without cross-validation, in a class of imbalanced problems from the bioinformatics domain. Ensemble-based scheme using the notion of diversity was also investigated. The experimental results favor the generation of neural network ensembles as these are able to produce good generalization ability and significant improvement compared to other single classifier methods [6].

Jiancheng Zhong et al. [35] presents 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 W 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. Liqi Li et. al [22] 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.

Muhammad Javed Iqbal et al. [16] 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.

A.Nisthana Parveen et al. [2] applied Principal Component Analysis (PCA) and Rough PCA feature selection approaches to discover discriminative features of Ecoli that will be the most adequate ones for classification.

Much as the above related work has been attempted, researchers in this piece of work aimed at obtaining a minimal dataset which is a representative of the most informative features of Ecoli and then applied a number of Classification techniques, Ten of them including ensembles. An insight is then done into a Comparison amongst all of these classifiers used to see the ones that perform better using a number of measures including Accuracy, Recall, Precision, F-measure, Hamming loss and Zero One loss, Time taken to produce desired results and Area under the curve (AUC).

3. METHODOLOGY

The E.coli dataset used was obtained from a Public Repository. The features and classes were further studied so that they can be put in a form where classification techniques could be applied.

3.1 The Dataset

This E.coli dataset used with protein localization sites was freely obtained from Lichman [21]. It includes 336 protein sequences labeled according to 8 classes including cp, im, pp, imU, om, omL, imL, and imS which are the localization sites. Table 1 illustrates the occurrence of classes for this dataset.

3.1.1 Preprocessing the data

The E.coli dataset [21] was in .data format. It was converted to xls format using Microsoft office Excel application. Data Cleaning was then performed where the attribute (Sequence Name) that was not needed for learning by machine learning classifiers was removed.
The dataset with all needed fields was converted to csv format using Microsoft office Excel application. The obtained csv E.coli dataset was converted into a Pandas Dataframe using Python Programming. The Dataframe was then converted to its Numpy-array representation, a format that machine learning classifiers required to be able to do the classifications.

3.2 Feature selection

Feature selection is a process which attempts to select more informative features. When features are many, they at times overpower main features for classification and in such a scenario, feature selection comes in to select the most informative ones and therefore improve the prediction accuracy and reduce the computational overhead of the classification algorithms [2]. Among the many feature selection techniques that exist, the following were applied on the E.coli dataset to extract only the most important features.

3.3.1 Tree-based Feature Selection

This method selects most important Features using forests of trees that has a Tree-based estimator to compute feature importance. This is used to discard irrelevant features. The sklearn.tree module and forest of trees in the sklearn.ensemble module [7] belonging to the Scikit-learn, an open source machine learning library used in Python programming were used when computing these features.

3.3.2 Recursive Feature Elimination (RFE)

Given an external estimator that assigns weights to features (e.g., the coefficients of a linear model), recursive feature elimination (RFE) selects features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and weights are assigned to each one of them. Features whose absolute weights are smallest are then pruned from the current set features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.
Obtaining the minimum dataset Resulting features from the above methods were combined to obtain a set that is a representative of the most informative features.

3.3 Classification of E.coli protein localization Sites

To predict protein localization sites, ten different classifiers including ensembles were applied. Ensembles include Random forest [26], Gradient boosting [25][13], Extreme Randomized Trees [25] and AdaBoost [28] [17]. Base classifiers are K-nearest Neighbors [12] [33] [14], Decision tree [19][25], Naive Bayes [14], Support vector classifier (Linear SVM) [11][18][24], Linear Discriminant Analysis [20] and RBF SVM [6]. The method used for splitting data set into training and testing was the k-fold cross-validation. The dataset was randomly split into 5 mutually exclusive subsets (folds) of equal size [23]. The implementation platform was Ipython with Pandas and Scikit-learn libraries.

3.4 Evaluation of Performance of the different Feature Selection and Classification methods

The choice made on which Classifier performs best was based on the results of Accuracy, Recall, Precision, F-measure, Receiver operating characteristic, Hamming loss and Zero One loss.
Accuracy was computed as;

$$Accuracy = \frac{TP + TN)}{(TP + FP + FN + TN)} \quad (1)$$

Recall was computed as;

$$Recall = \frac{(TP)}{(TP + FN)} \quad (2)$$

Positive Predictive Value (PPV) was computed as;

$$Recall = \frac{(TP)}{(TP + FP)} \quad (3)$$

Table 1. Description of the Dataset

Class	Label used (Abbreviation)	Number
Cytoplasm	cp	143
Inner membrane no signal sequence	im	77
Periplasm	pp	52
Inner membrane, uncleavable signal sequence	imU	35
Outer membrane non-lipoprotein	om	20
Outer membrane lipoprotein	omL	5
Inner membrane lipoprotein	imL	2
inner membrane, cleavable signal sequence	imS	2

A comparison of the true positive rate and true negative rate for the different classifiers was done by use of **Area under the curve (AUC) analysis**. Area under the curve of a receiver operating characteristic (ROC) curve is a way to reduce ROC performance to a single value representing expected performance [9]. Receiver operating characteristics graphs are useful for organizing classifiers and visualizing their performance. The hamming loss zero one loss was further computed to see the error rate produced by each of the classifiers.

4. RESULTS

This section presents the results obtained from Analysis of the data. The choice made on determining which of the classifiers performed best was based on the results of the AUC analysis and Confusion matrix. A comparison of the true positive rate and false positive rate for the different classifiers was done. A classifier that gives an AUC score of 1.0 has predicted perfectly. An area of 0.5 represents a worthless test and below 0.5 means the classifier is anti-correlated with the target.
Classification techniques were applied on the most informative features as obtained using feature selection methods.

4.1 Reduced Features as obtained by different methods

The most important features were obtained using a number of techniques Table 2. Figure 1 shows the Feature rankings to show how the most useful features were obtained using Tree-based feature selection method. The selected features had importance factor $>$ 0.1 in the used methods Table 4. The features obtained using both methods Table 2 were then combined to get a Reduced dataset Table 5 that was used in training and performing predictions by the classifiers.

4.2 Performance of Classifiers

This section presents the results obtained from predictions made by different classification methods. The performance of these methods is also evaluated and the best algorithm is then recommended.

Splitting data using Cross Validation

K-fold cross-validation was used in splitting the data set into training and testing. In k-fold cross- validation, a dataset D is randomly split into k mutually exclusive subsets the folds, $D_1, D_2, ..., D_k$ of approximately equal size. Cross validation used was 5.
The classifier is trained and tested k times each time $t \in 1, 2, .k$, it is trained on all $D_i; i = 1, ..., k$ with i not equal to t and tested on D_t. The cross validation estimate of accuracy is the overall number of correct classifications, divided by the number of instances in the dataset [23]. Cross validation was used to estimate the accuracy of the ten classifiers. The dataset was randomly partitioned into equally sized subsets with the proportion of the classes being equal in each partition.

Accuracy by each of the Classifiers

The dataset was first partitioned into 4 portions using Cross-Validation. Columns named 1, 2, 3 and 4 denote Partition 1, Partition 2, Partition 3 and Partition 4 respectively. The classifiers were then applied on each partition to obtain accuracy scores. The mean and standard deviations (std) were then computed for the Accuracy scores on each of the partitions as shown in Table 6.
From the Standard deviation results obtained, Linear SVM, Random Forest and KNN have the least values whereas RBF SVM, Naive Bayes and AdaBoost have the highest. A standard deviation close to 0 indicates that the data points tend to be very close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.

Results from PPV, F-measure, Recall Scores and losses for the Performance of Each Classifier Using 5-CV.

Under this section, the results of PPV, f-measure, and recall are given in Percentage. The Hamming loss and Zero one loss and the Time in seconds each technique takes to produce the results is given in Table 7. LSV represents linear SVM, DT Decision Tree, AD AdaBoost, EXT ExtraTree Classifier, RF Random Forest, NB Naive Bayes, GB Gradient Boosting, and RBF Radial Basis Function Support Vector Machine.
Results obtained indicate that Gradient Boosting is the best at performance 7 basing on Positive predictive value, recall and F-measure scores. It also produced the least error rate of 0.07 among all classifiers. Extra Tree Classifier, Random Forest, Naive Bayes and KNN also perform well. AdaBoost is the worst and produces the highest error rate of 0.27 (27%) in performance. DT took the least time (0.00 seconds) to execute followed by LDA, and Naive Bayes that took 0.01 Seconds.

Results obtained Using Area under the curve (AUC)

The ROC curve is created by plotting the fraction of true positives (TPR) against the false positives (FPR) at various threshold settings. When using AUC, Accuracy is measured by the area under ROC curve with area of 1 representing a perfect test; an area of 0.5 represents a worthless test. AUROCC has a range of 0 to 1.0, with accuracy tests of 0.90-1 regarded as excellent, 0.80-0.90 good, 0.70-0.80 fair, 0.60-0.70 poor and 0.50-0.60 fail. ROC Curve performance can be visualized on the diagram as in Figure 2
Table 8 corresponds to accuracy obtained by AUC on Ecoli's reduced dataset.

Table 2. Methods Showing most Informative features

No	Methods	Feature listed in order of their importance
1	Tree-based feature selection	alm1, mcg, alm2, gvh and aac
2	Recursive feature elimination (RFE)	mcg, gvh, alm1 and alm2

Table 3. Most Informative features.

Table 4. Showing most Informative features and their importance factors as obtained using Tree-based feature selection method

No	Feature	Factor
1	alm1	0.273316
2	mcg	0.208026
3	alm2	0.193493
4	gvh	0.168856
5	aac	0.126560
6	lip	0.026341
7	chg	0.003409

Results obtained Using Confusion Matrix

Confusion Matrix has been used to show the actual class labels in the vertical column and predicted class labels in the row across the top. It identifies misclassifications for each of the classifiers. In this piece of work, AdaBoost had majority of misclassifications for the different proteins as seen in Table 10. 30 imU proteins are incorrectly predicted to be localized to im, and 19 om proteins to be localized to pp. This shows the worst performance compared to the rest of Classifiers.

Gradient Boosting is seen to get most of the classifications right. Very few misclassifications are found. For example only 1 protein that localizes to cp is incorrectly predicted to localize to pp. No misclassifications are seen with omL and imL Proteins. More so, only 2 proteins that localize to pp are incorrectly predicted to localize to cp and im as seen in Table 11.

Comparison of results obtained in this work with those obtained by previous researchers

Table 9 shows results of Cross-validation as obtained by [14] on E.coli Dataset. Columns 1, 2, 3 and 4 represent accuracies obtained on the four partitions of the dataset.
Looking at results obtained on the same dataset [14] as in Table 9, KNN's accuracy is 86.3%, Decision Tree 80.36%, Naive Bayes 80.95% compared to 87%, 83% and 86% 9 for Decision Trees, Naive Bayes and KNN respectively obtained in this research work and thus an improvement is realized.

Furthermore, the Confusion matrix obtained in this work shows improvements in correctly classifying the Protein Localization Sites as compared to those obtained by others using KNN [14]. With Classifiers in this work, Only 1, 16, 7 were misclassified as compared to 2, 19, and 12 in cp, im and imU proteins respectively by other researchers [14].

5. CONCLUSION

A range of techniques have been used in attempt to correctly achieve prediction of the localization of proteins with majority of them responding positively. Extra Tree Classifier and Gradient Boosting are seen to be the best in performance followed by Random forest as seen from Precision, Recall and F-measure scores. Ensembles generally performed better than other classifiers with a score of 98% accuracy using AUC. This has been proved by past researchers as ensembles are less probable to misclassify unseen data samples than a single classifier [15].
However, AdaBoost was the worst at performance with an accuracy of 83%. Its Accuracy was still the worst at 71 % using Cross validation . The poor performance of AdoBoost observed is attributed to its limitations especially when applied to multi-class data problems as discussed by Tae-Hyun Kim et al. [32] which is the case with the problem researchers are addressing.

6. FUTURE WORK

—1. Training the classifiers on datasets for other living organisms that are especially responsible for waterborne diseases. This could help scientists involved in drug discoveries to discover drugs for diseases caused by such organisms

—2. Other feature selection methods apart from the ones used should be used to see if the same minimal dataset is obtained.

—3. Other data mining tools such as Tanagra with inbuilt classification techniques should be used to classify the obtained Ecoli minimal dataset so that results obtained are compared with the ones got in this research work.

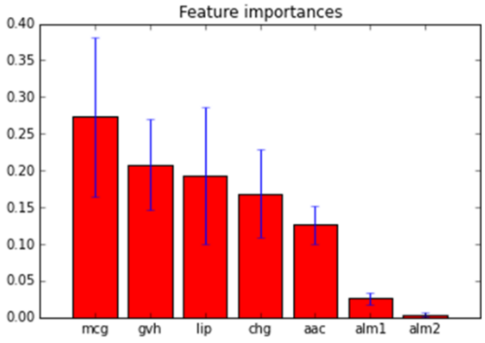

Fig. 1. Feature Importance Rankings

Table 5. Most informative Features Obtained from both methods

No	Attribute
1	alm1
2	mcg
3	alm2
4	gvh
5	aac

Table 6. Accuracy scores for the classifiers using Cross validation

No	Classifier	1	2	3	4	Mean	Std
1	KNN	90	87	86	86	87	1.639
2	Linear SVM	78	76	78	76	77	1.0
3	RBF SVM	68	78	73	74	73	3.562
4	Decision Trees	84	79	84	84	83	2.165
5	Random Forest	84	84	81	84	83	1.230
6	AdaBoost	70	76	70	67	71	3.269
7	Naive Bayes	87	90	84	81	86	3.354
8	LDA	90	86	88	91	89	1.920
9	ExtraTree Classifier	86	83	80	84	83	2.165

Table 7. Precision, Recall and F-measure scores for each of the Classifiers

Classifier's measure	KNN	LSV	RBF	DT	RF	LDA	AD	EXT	NB	GB
PPV	0.87	0.72	0.86	0.88	0.89	0.87	0.66	0.94	0.87	0.93
Recall	0.86	0.76	0.86	0.88	0.88	0.87	0.68	0.94	0.86	0.93
F-measure	0.87	0.82	0.87	0.88	0.88	0.86	0.73	0.94	0.86	0 0.93
Zero one loss	0.13	0.18	0.13	0.12	0.11	0.27	0.27	0.06	0.14	0.07
Hamming loss	0.13	0.18	0.13	0.12	0.11	0.27	0.27	0.06	0.14	0.07
Time (s)	0.02	0.02	0.03	0.00	0.04	0.01	0.13	0.04	0.01	1.11

Fig. 2. ROC curve

Table 8. Accuracy obtained by AUC on Ecoli's reduced dataset.

No	Classifier	Accuracy
1	KNN	0.98
2	Linear SVM	0.98
3	RBF SVM	0.99
4	Decision Trees	0.97
5	Random Forest	0.98
6	AdaBoost	0.83
7	Naive Bayes	0.96
8	LDA	0.98
9	ExtraTree Classifier	0.98
10	Gradient Boosting	0.98

Table 9. Comparison of results of Cross-validation

No	Classifier	1	2	3	4	Average	Std
1	KNN	89.28	95.24	84.52	76.19	86.31	8.04
2	Decision Trees	83.33	80.95	88.10	69.05	80.36	8.10
3	Naive Bayes	82.14	84.52	82.14	75.00	80.95	4.12

Table 10. Confusion Matrix for AdaBoost

Observed	Predicted								
cp (143)	127	0	0	0	0	0	0	16
im (77)	2	72	1	0	0	2	0	0
imL (2)	0	1	1	0	0	0	0	0
ims (2)	0	1	0	0	0	0	0	1
imU (35)	0	30	4	0	0	0	0	1
om (20)	0	0	0	0	0	1	0	19
omL (5)	0	0	0	0	0	1	0	4
pp (52)	3	1	0	0	0	3	0	45

Table 11. Confusion Matrix for Gradient Boosting

Observed	Predicted							
cp (143)	142	0	0	0	0	0	0	1
im (77)	1	70	1	0	5	0	0	0
imL (2)	0	0	2	0	0	0	0	0
ims (2)	0	1	0	0	0	0	0	1
imU (35)	1	4	2	0	28	0	0	0
om (20)	0	1	0	0	0	16	0	3
omL (5)	0	0	0	0	0	0	5	0
pp (52)	1	1	0	0	0	0	0	50

References

[1] Nerino Allocati, Michele Masulli, Mikhail F, and et al. Article: Escherichia coli in europe: An overview. *International Journal of Environmental Research and Public Health*, 10(12):6235–6254, 2013.

[2] A.Nisthana, H.Hannah Inbarani, and E.N. Sathish Kumar. Performance analysis of unsupervised feature selection methods, June 2013. https://arxiv.org/pdf/1306.1326.pdf.

[3] A D Aristoklis and M D George. Analysing the localisation sites of proteins through neural networks ensembles. *Neural ComputApplic*, 6(162), Jan 2006.

[4] P Trouiller J Pinel B Pcoul, P Chirac. Access to essential drugs in poor countries: a lost battle? *Journal of the American Medical Association*, 281(4):361–367, 1999.

[5] H Bouziane, B Messabih, and A Chouarfa. Isolation and antibiotic susceptibility of e. coli from urinary tract infections in a tertiary care hospital. *International Journal of Computer Theory and Engineering*, 5(4), 2013.

[6] H Chih-Wei, C Chih-Chung, and L Chih-Jen Lin. A practical guide to support vector classification. 2010.

[7] HT Debas, R Laxminarayan, and title = SE Straus'.

[8] V H Gunnar N Henrik E Olof, B Sren. Locating proteins in the cell using targetp, signalp and related tools. *Nature Protocols*, 2:953–971, April 2007.

[9] T. Fawcett. An introduction to roc analysis. *Pattern Recogn. Lett*, 27(8):861–874, 2006.

[10] D. Gould. Prevention and treatment of escherichia coli infections. *Nurs Stand*, 24:50–6, 2010. PubMed PMID: 20441035.

[11] S. R. Gunn. Support vector machines for classification and regression.

[12] HafidaBouziane, BelhadriMessabih, and AbdallahChouarfia. Meta-learning for escherichia coli bacteria patterns classification. 2012.

[13] J. He and B. Thiesson. Asymmetric gradient boosting with application to spam filtering. August 2007.

[14] P Horton and K Nakai. Better prediction of protein localization sites with the k nearest neighbours classifier. 1997.

[15] K-W Hsu. A theoretical analysis of why hybrid ensembles work. *Computational Intelligence and Neuroscience*, July 2017.

[16] MJ Iqbal, I Faye, and BB Samir. Efficient feature selection and classification of protein sequence data in bioinformatics. *The Scientific World Journal*, 2014.

[17] Z Hui H Trevor J Zhu, SaharonRosset. Multi-class adaboost. Jan 2006.

[18] S. Keerthi, O. Chapelle, and D. DeCoste. Building support vector machines with reduced classifier complexity. *Journal of Machine Learning Research*, 7:1493–1515, 2006.

[19] R.-H. Li and G. G. Belford. Instability of decision tree classification algorithms. pages 570–575, 2002.

[20] O Mitsunori Li Tao Li, Z Shenghuo. Using discriminant analysis for multi-class classification: an experimental investigation. *nowledge and Information Systems*, 10(4):453, 2006.

[21] M Lichman. Uci machine learning repository. 2013.

[22] L Liqi, Y Sanjiu, and X Weidong et.all. Prediction of bacterial protein subcellular localization by incorporating various features into chou's pseaac and a backward feature selection approach. *Biochimie*, 20(5):100–107, 2014.

[23] M. E. MacIntyre, B. G. Warner, and R. M. Slawson. Escherichia coli control in a surface flow treatment wetland. *Journal of Water and Health*, 4(2), 2006.

[24] Z Nina and W Lipo. A novel support vector machine with class-dependent features for biomedical data. October 2006.

[25] F. Pedregosa, G. Varoquaux, and A. Gramfort. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] title = Q Yanjun'.

[27] FS Brinkman S Rey, JL Gardy. Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. *BMC Genomics*, 6(162), 2005.

[28] R E. Schapire. A brief introduction to boosting. 1999.

[29] S.Sabir. Isolation and antibiotic susceptibility of e. coli from urinary tract infections in a tertiary care hospital. *Pakistan Journal of Medical Sciences*, 30(2):389–392, 2014.

[30] S Sterckx. Patents and access to drugs in developing countrimid: 1508637es: an ethical analysis. *Dev World Bioeth*, 4(1):58–75, 2004.

[31] T Akutsu T Tamura. Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition. *BMC Bioinformatics*, 8(466), Jan 2007.

[32] K Tae-Hyun, P Dong-Chul, and W Dong-Min. Multi-class classifier-based adaboost algorithm. *Springer*, 2011.

[33] Z. Voulgaris and G. D. Magoulas. Extensions of the k nearest neighbour methods for classification problems.

[34] C Yetian. Predicting the cellular localization sites of proteins using decision tree and neural networks. 2016.

[35] J Zhong, J Wang, and W Peng et.all. A feature selection method for prediction essential protein. *IEEE*, 20(5):491–499, 2015.