UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Classifying Virus Strain Using a Machine Learning Model Based on Subcellular Localization Data

Muhammad Izzat Kamaruddin[1], Rohayanti Hassan[2], Nabil Rayhan[3]
& Muhammad Luqman Mohd-Shafie[4]
Faculty of Computing
Universiti Teknologi Malaysia
81310 UTM Johor Bahru, Johor, Malaysia
Email: m.izzat@graduate.utm.my[1], rohayanti@utm.my[2], rayhan.nabil@graduate.utm.my[3],
muhammadluqman.ms@utm.my[4]

Submitted: 17/1/2024. Revised edition: 28/3/2024. Accepted: 12/5/2024. Published online: 31/5/2024
DOI: https://doi.org/10.11113/ijic.v14n1.460

***Abstract*—The topic of mRNA subcellular localization is very useful for further study. And one of the most significant reasons to study deep into this topic is to study mRNA functions. The location of the particular mRNA is very important, as well as its function. Localization of mRNA can be used for a variety of reasons. Therefore, several tools were developed to predict mRNA localization. Due to the various importance and functions of subcellular localization, further studies and research have been given significant attention by the researchers. Among all of the tools developed, some notable differences between those existing machine learning models are the methods implemented within the models. These methods give huge impacts on the outcomes of the prediction model. In this paper, the research focuses on analyzing the methodology and performance of mRNA subcellular localization prediction models.**

***Keywords*—mRNA, Subcellular, Localization, Machine Learning**

I. INTRODUCTION

Bioinformatics is an interdisciplinary field that develops methods and applies computational algorithms and tools to understand, visualize and analyze cellular data. In other words, it combines several fields such as biology, computer science, mathematics, and statistics to interpret large and complex biological datasets. The field of bioinformatics deals with database management and analysis, including DNA, RNA, and protein sequence data. Bioinformatics has experienced rapid growth and evolution in recent years, including the implementation of machine learning algorithms for further analysis and study of biological datasets.

A number of diseases and epidemics have plagued the world in recent decades, including Mweka ebola, Cholera, and Covid-19. This pandemic has crippled various sectors of the world, especially the economic and health sectors. As time has passed, many studies and research have been conducted in order to identify the right cure for the disease, thus resolving the pandemic. As Bioinformatics and the medical field developed and grew rapidly, various statistical and machine learning algorithms were developed and used to improve and increase the effectiveness and efficiency of healthcare. One of the key processes for a more detailed and deeper study of a particular disease is to trace and study subcellular localization and activity within the cell. Subcellular localization of mRNA is a biological process associated with this activity.

Currently, the topic of mRNA subcellular localization is one that is of great interest to study in depth. The study of the functions of mRNA is one of the most important reasons for going deeper into this topic. It is very important to understand not only the function of the mRNA but also the location of the particular mRNA localized. Hence, a few tools were developed to predict mRNA localization. The publicly accessible prediction techniques vary primarily in four ways that are relevant to the user which are the underlying biological purpose, the computational approach utilized, the localization coverage, and reliability [1].

In this study, we analyze the performance of two Machine Learning classification models, namely Naive Bayes and K-Nearest Neighbor classifiers. In addition, this project also aims to evaluate the performance of the Machine Learning classifier using classification evaluation metrics.

II. RELATED WORKS

In the process of protein synthesis, messenger RNA (mRNA) plays an important role. MRNA carries the coding segment generated from deoxyribonucleic acid transcription (DNA) which is carried out by RNA Polymerase. There are four main factors that influence the localization of mRNA after it is transcribed, which are vectorial export from nuclei, localized degradation protection, polarized active transport on the cytoskeleton, and localized anchorage. It will be possible to control protein production quantitatively and spatially by subcellular localization [2]. In particular, this process provides a cost-effective method of protein localization by transporting messenger RNA to the subcellular location where the protein is required, followed by on-site translation.

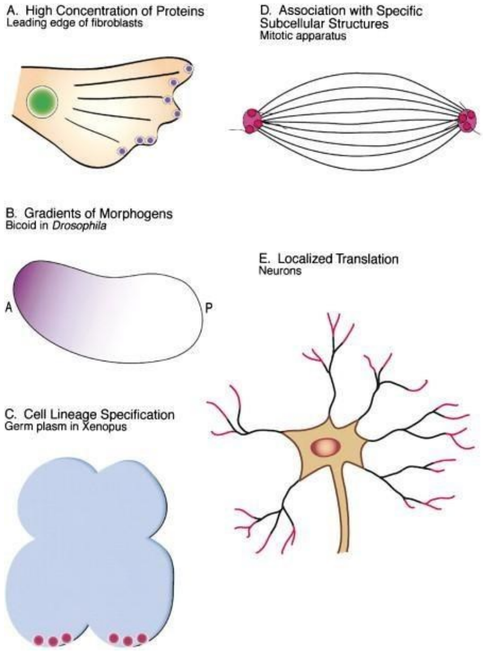
A. High Concentration of Proteins
Leading edge of fibroblasts
D. Association with Specific Subcellular Structures
Mitotic apparatus
B. Gradients of Morphogens
Bicoid in *Drosophila*
A
P
E. Localized Translation
Neurons
C. Cell Lineage Specification
Germ plasm in *Xenopus*

Fig. 1. Different roles of mRNA Localization (Malgorzata *et al*., 2002)

Subcellular localization of mRNA can be triggered by a variety of factors and mechanisms. In order to produce a high concentration of proteins in a particular cell area, a first factor is to promote the development of high protein concentrations there [3]. In Fig. 1 (A), the localization of actin protein to a particular region of the cell may lead to an increase in the concentration of actin protein at the front edge of the cell. Similarly, mRNA localization is essential for segregating specific mRNAs to specific subcellular locations (D) as shown in Fig. 1. As an example, Cyclin mRNA is found near the poles of mitotic spindle fibers [4]. Some mRNA localization processes are directed to specific subcellular locations in order to limit translation at specific locations, while others are not. MRNA localization is a follow-up of transcription within The majority of mRNA leaves the nucleus via nuclear pores and enters the cytoplasm for translation. Nevertheless, some mRNAs are transcripts with specific destinations and target a specific region within the cell. When the mRNA reaches its final destination, it is translated. Within the cell, cis-acting elements are responsible for localization. As a result of these signals, trans-acting factors will be directed to bind to the mRNA. Through the interaction of these two molecules, mRNA structures are altered, resulting in the folding of the mRNA into a particular spatial configuration which then facilitates protein association. Proteins within a subcellular compartment were associated with its physiological and metabolic function [4]. Predicting the subcellular location of mRNA after it has been localized may provide information about the function of the gene from which the mRNA was produced. The tool that can predict the correct subcellular location of transcripts may therefore contribute to the understanding of gene expression regulation. The use of computational predictors is one of the most effective methods of predicting the subcellular localization of mRNA. A variety of computer techniques have been used to identify and track the subcellular location of single RNAs in live cells in recent years.

In 2019, RNATracker was introduced as the first tool for predicting mRNA subcellular localization. Using the recurrent neural network (RNN) method, this computational mRNA localization predictor makes predictions based on raw mRNA input sequences. By masking 100 nucleotides of the sequence at a time, RNATracker can also identify possible zip codes [2]. To address the rapid growth in RNA localization data, Zhang *et al*. [5] introduced 'iLoc-mRNA', another tool for predicting mRNA localization. LOS-mRNA is a machine-learning approach to predicting mRNA subcellular localization in Homo Sapiens. This mRNA prediction tool was developed using a support vector machine (SVM) deep learning approach and a collection of optimally selected features. To predict the localization of eukaryotic mRNA in five subcellular compartments, mRNALoc uses the SVM method based on pseudo-K-tuple nucleotide composition characteristics. The most recent computational predictor, DM3LOC, incorporates a deep learning technique with a multi-head self-attention approach to predict mRNA subcellular localization.

Machine Learning classifiers are one of the most effective approaches in the field of class classification. It's a branch of science that focuses on ways for computer systems to enhance their performance by learning (or changing their behavior) from previous data instances. During the learning process, structural patterns in the provided dataset ("training set") are developed. When presented with data that has been classified in an unknown manner ("test set"), these patterns are then used as the basis for making predictions. A number of machine learning algorithms have been used to predict the subcellular localization of proteins in previous studies. Since proteins located in certain cellular compartments have several characteristics, certain machine learning algorithms have been used to predict their subcellular localization. This study utilized a number of machine learning approaches. Among these were J48, a decision tree algorithm, SVM, Multi-Layer Perceptron (MLP), a Neural Network implementation, and Naive Bayes. Machine Learning employs three types of amino acid sequence characteristics: composition, transition, and distribution. Several of these characteristics have been effectively utilized in machine learning techniques to predict protein secondary structure and subcellular location.

A. Naive Bayes Classifier

A naive Bayes classifier is one of the most used practical Bayesian learning algorithms. Naive Bayes is also one of the good examples of supervised learning algorithms that have shown to be not only easy, but also quick, accurate, and dependable in their use. Naive Bayes has been successfully used in many works, especially with Natural Language Processing (NLP) problems [6].

Naive Bayes is a probabilistic machine learning technique that is based on the Bayes Theorem and is utilized in a broad range of classification problems, such as image classification and text classification. Bayesian classifiers determine if a particular attribute value has an influence on a certain class based on the values of the other attributes. Class conditional independence is the term used to describe this assumption. It is designed to make the calculation involved as simple as possible, and as a result, it is seen as "naive."

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \qquad (1)$$

where,

A|B	=	Probability of A occurring given evidence B has already occurred
B|A	=	Probability of B occurring given evidence A has already occurred
A	=	Probability of A occurring
B	=	Probability of B occurring

B. K Nearest Neighbour Classifier

In Machine Learning, non-parametric approaches are referred to as instance-based or memory-based algorithms. As a result, this algorithm does not contain any formulas or calculations. The K-Nearest Neighbour algorithm tends to store the training instances in a lookup table and interpolate them. The K-Nearest Neighbor algorithm is one of the most fundamental nonparametric methods. This is a simple deep learning algorithm that is capable of handling both classification and regression problems. This algorithm differs from other classification algorithms that implement artificial neural networks by requiring the training set to include both positive and negative cases. It works by calculating the distances between a query and all of the instances in the data, selecting the number of examples (K) that are closest to the query, and then selecting the most frequently occurring label (in classification) or averaging the labels (in regression) [7]. Cross-validation is generally used to determine the appropriate value of k for reducing noisy points within the training data set.

C. Classifier Performance Measurement

Classifier models were developed and programmed to learn from an infinite number of different data sets, referred to as training sets. Classifier models were developed experimentally using a variety of multi-set data, also known as test data. It is important to note that the success of the models when applied to the test datasets serves as a proxy for the success of the model when applied to other multi-sets of data. Machine learning hypotheses must be tested, which is why every model must be evaluated. An assessment of a classifier may be based on a numerical measure, such as accuracy, or a graphical representation, such as a receiver operating characteristic curve (ROC). There are advantages and disadvantages to both of these methods.

III. METHODOLOGY

A series of phases and activities has been designed to ensure that a methodical analysis of the Machine Learning prediction models is conducted. There are six phases in the research framework, with the first focusing on research planning and preliminary research. In the second step, data is prepared, and in the third phase, input datasets are prepared, including data cleaning and data processing. A fourth phase of the study focuses on the development and evaluation of Machine Learning classification algorithms, and a fifth phase focuses on the analysis of model performance and outcomes. The final phase of the process is the evaluation and discussion. This study uses the performance measurement of the classification technique as a measure to compare the existing models for predicting mRNA subcellular distribution. The flow of the experimental design is depicted in Fig. 2.

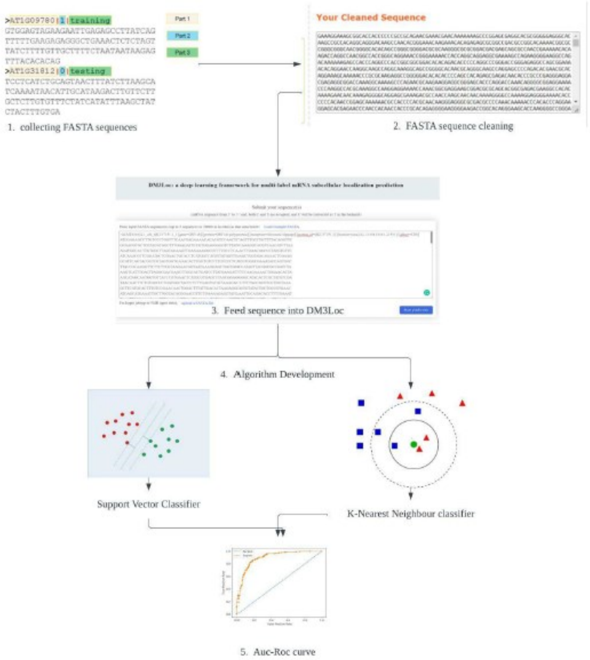
Part 1
Part 2
Part 3
1. collecting FASTA sequences
Your Cleaned Sequence
2. FASTA sequence cleaning
3. Feed sequence into DM3Loc
4. Algorithm Development
Support Vector Classifier
K-Nearest Neighbour classifier
5. Auc-Roc curve

Fig. 2. The experimental design

A. Data Collection and Preparation

The main idea of the datasets in this research is to get the RNA sequences of viruses that ever resulted in epidemics or outbreaks. Different species of the virus were selected such as Human Coronavirus, Porcine respiratory coronavirus, and Middle East respiratory syndrome-related coronavirus (MERS-CoV). For every species of virus selected, a primary sequence of the virus RNA was obtained from the NCBI database.

To train a good model, a huge number of datasets are required [8]. After obtaining the primary sequences for every species of virus, the next step is to increase the size of the datasets. In this phase, there are a few steps included as shown in Fig. 3.

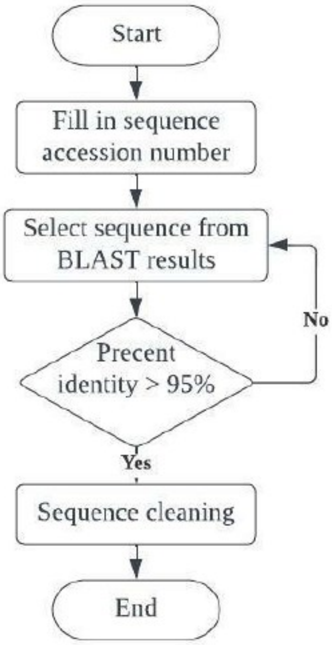
Start
Fill in sequence accession number
Select sequence from BLAST results
Precent identity > 95%
No
Yes
Sequence cleaning
End

Fig. 3. The process flow of data collection

The data preparation process continues with data cleaning. This process was carried out by using the online sequence cleaner, The Bio-Web. Through this online tool, the sequences obtained in the previous step were cleaned from unwanted text marks, non-templated nucleotides, and spaces. This is to ensure that the sequences that were used in this research are cleaned from data noise thus contributing to good results. This process is also very important to get rid of redundant sequences. When one or more similar/homologous sequences are present in the same set of data, it is called redundancy in a collection of data sequences.

B. Data Processing

The prediction tool that will be used in this phase was fed with the datasets prepared in the previous phase. As stated in the early chapter, the main prediction tool that will be used in this experiment is DM3Loc. The main goal of this process is to obtain the prediction score of every sequence. The datasets that were obtained in the previous phase, were fed into the DM3Loc predictor. The results derived from the tool were recorded and tabulated for further analysis and evaluation process. As different tools were built with different mechanisms and algorithms, hence the results produced are also different in form of prediction scores and output numbers. As DM3Loc was designed with a deep-learning method of multi-head self-attention, the RNA subcellular localization predictions will be producing results in multiple compartments. The model accepts mRNA sequences at variant lengths. It encoded the mRNA sequences using one-hot encoding for four types of nucleotides where T and U share the same coding [9].

C. Algorithm Development

The phase dedicated to algorithm development focused primarily on two key machine learning classifiers: the Naive Bayes classifier and the K-Nearest Neighbor classifier. In both cases, supervised classifiers are used to assign a new object to a specific class based on its attributes and a training dataset. Several activities were conducted during this phase, including the representation of datasets, the development of classification models, model training, model testing, generating confusion matrices for predictions, and training the proposed classifier model. We obtained critical classification results from this phase, which were then analyzed and discussed in the following phase of our research.

It was the primary goal of the training process to minimize the disparity between the prediction vector and the true label vector when addressing the classification problem. This phase involved meticulously developing and rigorously testing two classifier models using the sequences and datasets collected during the data collection process. The main programming platform for this phase was Jupyter Notebook, a web-based Python development environment.

In order to train the model, the datasets were first divided into training and testing subsets. A total of 80 percent of the data was allocated for training and 20 percent for testing, resulting in a real sample size of 401 and 101, respectively. Data partitioning was used to train the models with multiple datasets from the testing subset in order to make predictions based on the training datasets. Following this, the models were rigorously tested using the designated testing datasets.

1) Naive Bayes Classifier

The Gaussian Naive Bayes model is the most suitable model for continuous values of datasets and features that follows normal distributions. By using the *sk.learn* library, the Gaussian Naive Bayes algorithm was imported and then the model will proceed for model training and testing by using the training and testing datasets accordingly.

2) K Nearest Neighbour Classifier

The next step is to develop another classification model that was used in this experiment which is K Nearest Neighbour (KNN) classifier. KNN classification is the process of categorizing a given collection of data into several categories. KNN classification may be used for both structured and unstructured data types. The ideal choice of this k-value is greatly dependent on the data. In general, a larger k-value reduces the impacts of noise, but it also blurs the classification boundaries. The KNN algorithm was imported from *sklearn. neighbor* library and the experimental datasets were retrieved. The next process is model training and testing by using the training and testing datasets.

D. Evaluation of Classification Models

As the classification model was developed and trained using the datasets, further studies were conducted on its performance.

To evaluate the prediction models, a number of important metrics were highlighted, including precision, recall, and F1 scores. The recall metric is used to assess the model's ability to identify positive samples. A precision metric indicates the accuracy of a positive prediction made by a model, while a recall metric is calculated by dividing the total number of positive samples by the number of positive samples accurately identified as positive. The higher the recall, the greater the number of positive samples found. As one of the most important evaluation metrics, the F1 score integrates two measures, accuracy and recall, in order to summarize a model's predictive ability. These metrics are performed by importing the "classification_report" from the *sklearn.metrics* library. In this experiment, Naive Bayes and K Nearest Neighbor classification models were developed, trained, and tested accordingly. Different algorithms produced different outcomes and results that reflect the performance of the classification model.

IV. RESULT AND DISCUSSIONS

This research uses the data of sub-cellular localization prediction scores which were generated manually from some sequential processes. These prediction scores datasets were derived from the raw mRNA sequence of some selected species of virus which were obtained from the NCBI database. A total number of 503 mRNA sequences were used which then generated 503 rows of prediction scores. The generated datasets consist of prediction scores of different sub-cellular compartments including the nucleus, mitochondrion, endoplasmic reticulum, and others. Fig. 4 shows the presentation of the total datasets used in this research.

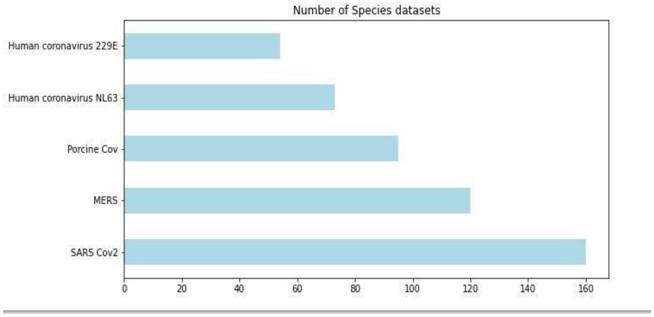

Fig. 4. Datasets presentation

As stated in the previous chapter, two prediction algorithms were developed and tested in this research, which are the Naive Bayes classification model and the K Nearest Neighbour classification model. These algorithms produce different outcomes and results that represent the performance of the classification model. Fig. 4 shows the prediction scores of Naive Bayes and K Nearest Neighbour classifiers on the testing and training datasets.

In this research, the evaluation of the prediction model is analyzed based on the accuracy, precision, recall, and f1 - score. These performance measurement metrics were derived from the confusion matrix which was generated by the sklearn.metric library. As mentioned in the previous chapters, the confusion matrix consists of 4 categories of results which are True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Based on the confusion matrix, the performance measurement metrics were calculated. The equation for the performance measurement variables is explained below:

$$ACCURACY = \frac{TP + TN}{TP + TN = FP + FN} \qquad (2)$$

In this experiment, the accuracy score shows how accurate the prediction results produced by the classification models are compared to the actual data. Table 1 shows the comparison of accuracy scores by the classification models. Based on the table shown, the KNN classifier model shows a better accuracy score compared to the Naive Bayes classifier model.

TABLE 1. ACCURACY SCORE

Accuracy		
Classifier	Naive Bayes	KNN
	0.95	0.91

The precision score is a measure of the accuracy of a model's positive predictions. The precision of a prediction is calculated by dividing the number of true positives by the total number of positive predictions. The precision scores from both classifier models are shown in Table 2. The results of the KNN classifier model are more consistent and produce a higher precision score, as shown in the table.

TABLE 2. COMPARISON OF MODEL PRECISION

Classifier	**Precision**				
	Human Corona virus 229E	Human Coronavirus NL63	MERS	Procine Cov	SARS CoV2
Naïve Bayes	0.94	0.76	0.94	0.40	0.25
KNN	0.82	0.83	0.96	0.87	1

The recall refers to the percentage of Positive samples that are correctly classified as Positive as compared with the total number of Positive samples. Recall is a measure of the model's ability to identify positive samples. The higher the recall, the greater the number of positive samples detected. Table 3 presents a comparison of recall scores for both classifier models. KNN classifier model produces more consistent results and produces higher recall scores in the table shown.

TABLE 3. COMPARISON OF RECALL SCORE

Classifier	Recall				
	Human Coronav irus 229E	Human Coronav irus NL63	MERS	Procine Cov	SARS CoV2
Naïve Bayes	1	0.93	0.71	1	0.03
KNN	0.90	1	0.92	1	0.86

The F1-score is one of the most important evaluation metrics. Through the combination of accuracy and recall, which are normally opposing measurements, it provides a concise summary of the prediction performance of a model. In Table 4, the f1-scores are compared between the two classifier models. KNN classifier models produce better and more consistent f1-scores than Naive Bayes classifier models. Considering that the f1-score takes into account both precision and recall scores, a higher precision and recall score will result in a higher f1-score.

TABLE 4. COMPARISON OF F1-SCORE

Classifier	**F1-score**				
	Human Coronavirus 229E	Human Coronavirus NL63	MERS	Procine Cov	SARS CoV2
Naïve Bayes	0.97	0.84	0.81	0.57	0.06
KNN	0.86	0.91	0.94	0.93	0.93

However, the accuracy score of these classification models was taken from 5 testing processes. This is due to the inconsistency of the accuracy score produced by the models. There are several factors that may contribute to the occurrence of the problems. One of the contributing factors is the imbalance between the classes of datasets. In order to classify class imbalanced datasets, it is necessary to determine the most appropriate performance metrics to be used. In previous work, it has been demonstrated that imbalance can have a significant impact on the value and meaning of accuracy as well as certain other well-known performance metrics [10].

Besides evaluating classification models using the performance measurement metrics discussed in the previous topic, this study also applies other performance measurement techniques, such as Receiver Operating Characteristics and Area Under Graph. These two techniques are considered to be essential for measuring the performance of classification models in the early chapters of this research. Among the best performance measurement metrics for classification experiments at various threshold values is the AUC-ROC curve. ROC is a probability curve that represents the degree of separability. AUC represents the degree of separability. In other words, it indicates how well the model is able to distinguish between classes. In general, the higher the AUC score, the more accurate the model is in classifying 0 classes as 0 and 1 classes as 1. For AUC values between 0.9-1, excellent, good, fair, poor for AUC values between 0.7-0.8, and failed for AUC values between 0.5-0.6 [11].

Both models produced ROC curves and accuracy scores that were not satisfactory and convincing, despite good accuracy scores. A number of factors that may contribute to the poor AUC score have been identified. There will be a significant impact on the ROC curve and the AUC score due to the imbalanced class datasets. The ROC AUC is sensitive to class imbalance in that when there is a minority class, the definition of data as positive will have a significant impact on the AUC value [12].

The second factor that may contribute to the poor AUC score and ROC curve is the uncleansed data during the data conversion process. The classification results, which are categorical data, must be converted into numerical data in order to construct and plot the ROC curve. As a result, conflict may occur during data conversion, resulting in a poor ROC curve and AUC value.

V. CONCLUSION

This study generated mRNA sequence datasets for five different virus species, including Human Coronavirus 229E, Human Coronavirus NL63, MERS, Procaine Cov, and SARS CoV2, based on sequences obtained from the NCBI GenBank. Datasets were prepared using BLAST analysis to identify similar sequences, followed by data cleaning to remove unnecessary elements. To obtain prediction scores for each sequence, the cleaned mRNA sequences were processed using the DM3Loc sequence predictor. Using these datasets, two classifier models, Naive Bayes and K Nearest Neighbor, were developed and tested using Jupyter Notebook. To evaluate the performance of the model, metrics such as precision, recall, and F1-score were used, and the results were presented graphically. According to the study, mRNA classification models are still in their infancy and their accuracy depends on the quality of available mRNA sequences, indicating that advancements in techniques and a wide variety of datasets will enhance future research into the subcellular localization of mRNAs.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education, Malaysia (MOHE) under the Fundamental Research Grant Scheme (FRGS), FRGS/1/2023/ICT02/UTM/02/8. This work also was supported/funded by Universiti Teknologi Malaysia under UTM Fundamental Research Grant (UTMFR): Q.J130000.3851.21H94.

REFERENCES

[1] Dönnes, P., Höglund, A. (2004). Predicting protein subcellular localization: past, present, and future. *Genomics Proteomics Bioinformatics, 2*(4), 209-15. Doi: 10.1016/s1672-0229(04)02027-3. PMID: 15901249; PMCID: PMC5187447.

[2] Zichao Yan, Eric Lécuyer, Mathieu Blanchette. (2019). Prediction of mRNA subcellular localization using deep recurrent neural networks. *Bioinformatics*, *35*(14), i333-i342.

[3] Aakalu, G., Smith, W. B., Nguyen, N., Jiang, C., Schuman, E. M. (2001). Dynamic visualization of local protein synthesis in hippocampal neurons. *Neuron*, *30*, 489-502

[4] Jing Li, Lichao Zhang, Shida He, Fei Guo, Quan Zou. (2021). SubLocEP: A novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning, *Briefings in Bioinformatics*, *22*(5).

[5] Zhang, Z. Y., Yang, Y. H., Ding, H., Wang, D., Chen, W., & Lin, H. (2021). Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Briefings in Bioinformatics, 22*(1), 526-535. Doi: 10.1093/bib/bbz177.

[6] Islam, Mohammed J & Wu, Q. M. Jonathan & Ahmadi, Majid & Sid-Ahmed, Maher. (2010). Investigating the performance of Naive- Bayes Classifiers and K- Nearest Neighbor Classifiers. *JCIT. 5*, 133-137. Doi: 10.1109/ICCIT.2007.148.

[7] Amendolia, S. R. & Cossu, G. & Ganadu, M. L. & Golosio, Bruno & Masala, Giovanni & Mura, G. M. (2003). A

comparative study of K-Nearest Neighbour. Support vector machine and multi-layer perceptron for Thalassemia screening. *Chemometrics and Intelligent Laboratory Systems. 69*, 13-20. Doi: 10.1016/S0169-7439(03)00094-7.

[8] Malgorzata Kloc and N. Ruth Zearfoss and Laurence D. Etkin. (2002). Mechanisms of Subcellular mRNA Localization Cell. *108*(4), 533-544. Doi: 10.1016/S0092-8674(02)006.

[9] Zou, Q., Xing, P., Wei, L., & Liu, B. (2019). Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *Rna, 25*(2), 205-218. Doi: 10.1261/rna.069112.118.

[10] Luque, A., Carrasco, A., Martín, A., & de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition, 91*, 216-231. Doi: 10.1016/j.patcog.2019.02.023.

[11] Metz, C. E. (1978). Basic principles of ROC analysis. Seminars in nuclear medicine, *8*(4), 283-298. Doi: 10.1016/S0001-2998(78)80014-2.

[12] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine learning,* 233-240. 10.1145/1143844.1143874.