DNACLASSIFICATION FOR DETECTION OF E. COLIVIRUS INFECTION

V. Lakshmi Chetana[1], Assistant Professor, Department of Computer Science Engineering, DVR & Dr.HSMIC College of Technology (Autonomous), India.
Email: lakshmichetana@mictech.ac.in
S. Girish Chandra[2], Assistant Professor, Department of Computer Science Engineering, DVR & Dr.HSMIC College of Technology (Autonomous), India.
Email: interview.girishchandra@gmail.com
M. Kavya sree[3], UG Student, Department of Computer Science and Engineering, DVR & Dr.HSMIC College of Technology (Autonomous), India.
Email: kavyamalapati2132@gmail.com
B. Sai Sri[4], UG Student, Department of Computer Science and Engineering, DVR & Dr.HSMIC College of Technology (Autonomous), India. Email: bollasaisri@gmail.com
P. Maha Lakshmi Manogna[5], UG Student, Department of Computer Science and Engineering, DVR & Dr.HSMIC College of Technology (Autonomous), India.
Email: manognamanupolisetti@gmail.com
E. Lokesh Rama Swami[6], UG Student, Department of Computer Science and Engineering, DVR & Dr.HSMIC College of Technology (Autonomous), India.
Email: erralokesh4@gmail.com

ABSTRACT
E. coli (Escherichia coli) is a type of bacteria that is commonly found in the intestines of humans and animals. Most strains of E. coli are harmless, but some strains can cause illness. An E. coli infection occurs when a person ingests food or water that is contaminated with harmful strains of the bacteria. To overcome the previous problem, we present an efficient system for detecting the presence of the E. coli virus in a DNA sample. An MLP classifier model has been developed using a DNA dataset containing four types of DNA molecules (A, C, G, and T). The dataset is used to train the model, which then is used to classify the DNA sample into E. coli virus or not. The model has been tested and proven to be accurate in identifying the presence or absence of the virus.To make the system easily accessible, a website has been developed using Django Framework. The website allows users to input their DNA samples and get the results of the classification in real-time. Furthermore, the website has been designed with user-friendly features such as a clean interface and simple workflows.The proposed system offers an efficient and accurate way of identifying the presence of the E. coli virus in a DNA sample. The results of the MLP classifier model are used to provide the classification results on the website. The system is easy to use and provides quick results, making it a useful tool for researchers and medical professionals. The system is also useful for educational purposes, allowing students to learn about the technique of virus identification in a DNA sample.
Keywords: E. coli Virus, Deep Learning, Virus classification

1INTRODUCTION Escherichia coli, sometimes known as E. coli, is a species of bacteria that lives in both human and animal intestines. It is a gram-negative, rod-shaped bacterium that plays a significant role in the microbial ecosystem that naturally exists in the human gut. Even while the majority of E. coli strains are helpful to humans and even harmless, some of them can nevertheless infect and sicken people.Contaminated food, drink, person-to-person contact, or exposure to animal excrement are all ways that E. coli can spread. Many symptoms, such as diarrhoea, abdominal pain, fever, and dehydration, can be brought on by infections. In severe cases, renal failure brought on by E. coli infections can be fatal.There are various varieties of E. coli, and each has a unique combination of traits and the capacity to spread disease. Enteric and diarrheal illnesses are brought on by some strains of E. coli known as diarrheagenic E. coli (DEC). Enterohemorrhagic E. coli (EHEC), enteropathogenic E. coli (EPEC), enterotoxigenic E. coli (ETEC), enteroaggregative E. coli (EAEC), and enteroinvasive E. coli are the five major pathotypes that these strains fall under (EIEC).n general, even though E. coli

 DOI: 10.36893.JK.2023.V13I04N16.0033-0038 Copyright @ 2023 Author

can cause illness and infection, good hygiene habits like frequent hand washing and meticulous food preparation can assist to avoid transmission and lower the risk of infection.

2 LITERATURE SURVEY ON E. COLI

There has been a great deal of research on E. coli over the years, both to better understand the bacteria itself and to develop treatments for E. coli infections. Some areas of research on E. coli include:**Molecular biology**: This field of study looks at the genetics and biochemistry of E. coli, and how the bacteria interact with its environment.**Antibiotic resistance**: There has been concern about the rise of antibiotic-resistant strains of E. coli, and researchers are looking for new ways to combat these infections.**Vaccine development**: There are currently no vaccines for E. coli infections, but researchers are working on developing vaccines that could prevent infections caused by certain strains of the bacteria.**Food safety**: E. coli can cause foodborne illnesses, so researchersare looking for ways to prevent the contamination of food products.**Clinical treatments**: Researchers are also investigating new treatments for E. coli infections, such as new antibiotics and other drugs that could help combat the bacteria.

3.METHODOLOGY AND OVERVIEW

To create a successful deep-learning model, it is necessary to go through the following steps:

3 Collect Data
1. Explore and Pre-process
2. Model Architecture
3. Train and evaluate the model
4. Deploy

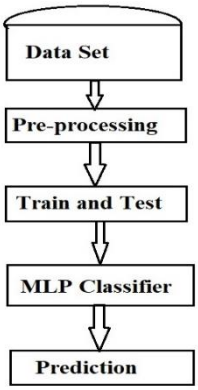

Fig 1. MODEL ARCHITECTURE

Collect Data

Data gathering is the first step of any Machine Learning or Deep Learning project. Without proper data, no model can give accurate results. Data can be gathered from various sources like websites, databases, surveys, etc. The data should be properly cleaned and pre-processed before training the model. It is important to select the right features for the model which will help in better predictions.

2. Explore and Pre-process

Once the data is collected, it is important to explore it and understand its characteristics. This helps in identifying data or any outliers which can affect the accuracy of the model. The data should be pre-processed accordingly such as normalizing, scaling, removing the outliers, etc. After pre-processing the data, the features should be selected and the data should be split into train and test sets

3. Choose a Model Architecture

The model architecture is an important part of any Deep Learning project, It should be chosen carefully based on the type and size of the data. The model should be capable of handling the data and giving accurate results. There is a variety of architecture available. The choice of architecture will depend on the problem that needs to be solved and the data that is available.

4. Train and Evaluate the model

 DOI: 10.36893.JK.2023.V13I04N16.0033-0038 Copyright @ 2023 Author

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

After the model architecture is chosen, the model should be trained by using various techniques like batch training, cross-validation, etc. The model should be evaluated by using different metrics such as confusion matrix, accuracy, precision, recall, etc. This will help in understanding the performance of the model and can be used to further improve the model.

5. **Deploy Model**

Finally, the model can be deployed in production. This involves setting up the infrastructure for the model, such as a web server, and making sure that the model can handle the traffic that comes through it. Once this is done, the model can be used to make predictions or generate insights from thedata

DATA PREPROCESSING

4.1 ONE HOT ENCODING

Categorical variables typically contain string values, which most machine learning algorithms cannot process. To use these variables in the algorithm, the strings must be replaced with numerical values through a process known as categorical variable encoding. This allows the algorithms to interpret the values and use them in the model.One Hot encoding is a popular method of categorical variable encoding. It is simple to implement and does not introduce additional bias into the model. This process involves creating a new dummy binary variable that is then categorized as the original categorical variable. The dummy variable is then assigned a value of 0 or 1 depending on whether they are present or absent in the data.For instance, if the DNA sequence is A, C, G, and T, then the one-hot encoding of the sequence would be [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0], and [0, 0, 0, 1], respectively

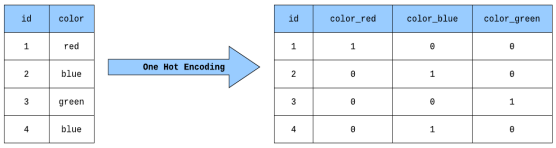

Fig 2. EXAMPLE OF ONE-HOT ENCODING

This process is useful because it reduces the dimensionality of the data while still retaining the information present in the original variable.

ARCHITECTURES

In deep learning, an architecture is a combination of layers and parameters that are connected to create a neural network. The architecture specifies how data is passed through each layer. Different architectures can be used to solve different tasks, and in this project, an MLP Classifier architecture is being used for classification purposes.

5.1 MLP CLASSIFIER

MLP stands for Multilayer Perceptron, which is a type of artificial neural network. It consists of an input layer, one or more hidden layers, and an output layer. Each layer is made up of interconnected

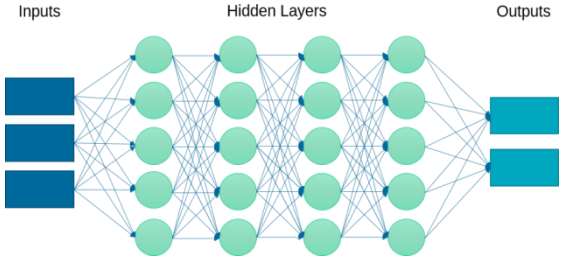

Fig 3. MLP CLASSIFIER

'nodes', which contain an activation function that determines the output of the node. The connections between nodes are represented by weights, which can be adjusted based on the output of the network. The architecture of an MLP is typically feedforward, meaning that the information flows in one direction from the input layer to the output layer, with no cycles or loops.

5.2 TRAIN AND TEST SETS

DOI: 10.36893.JK.2023.V13I04N16.0033-0038
Copyright @ 2023 Author

It is essential to divide the original dataset into the training dataset and the test dataset for deep learning applications. The test dataset is used to assess the model's performance and its capacity to generalize to new or unexplored data, whereas the training dataset is used to train the model. To accurately evaluate the model's accuracy and usefulness, this approach is required.With the train test split function from the sklearn library, we can divide the dataset into training and testing sets. The four variables that this function accepts are x train, x test, y train, and y test. The training data's features and dependent variables are represented, respectively, by the variables x train and y train. The independent variables and features for the testing data are represented, respectively, by the variables x test and y test.We supply four parameters to the function train test split(). Data arrays that we want to divide into training and testing sets make up the first two parameters. The size of the test set is specified by the third option, test size. Depending on the preferred ratio of training to testing data, we can choose the test size to be 0.5, 0.3, or 0.2.In order to guarantee that we always obtain the same result when we split the data, the final option, random state, is used to create a seed for the random generator.In general, creating accurate and efficient deep learning models requires dividing the dataset into training and testing sets. It enables us to assess the model's effectiveness and confirm that it is capable of generalizing well to fresh and untested data.

5.3 BUILDING THE MLP CLASSIFIER

Finally, we will build the Multi-layer Perceptron classifier.

- hidden layer sizes: With this option, we can choose how many layers and nodes the Neural Network Classifier should have. The number of nodes at each point is represented by each member in the tuple, where I am the tuple's index. Thus, the tuple's length denotes the total

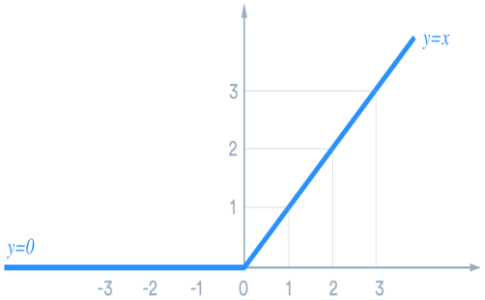

Fig4.GRAPHOF ACTIVATION FUNCTION

- max iter: This variable represents the number of epochs.
- Activation: The process that makes the concealed layers active.
- Solver: This parameter describes the algorithm for balancing weights among nodes.
- random state: This parameter lets you specify a set of seeds to get the same outcomes again.

We can now provide the data to the neural network for training after initialising it.

5.4 ACTIVATION FUNCTION: ReLU

The output of a neural network is determined by a mathematical formula called an activation function. It determines whether or not a neuron should be engaged by mapping the input signal to an output signal. The weighted sum of all the inputs from the preceding layers to the current layer is subjected to a nonlinear transformation.Graphical Representation of ReLU function: f(x)=max (0, x)
ReLU (Rectified Linear Unit) is a type of activation function used in neural networks. It is a non-linear function that is used to transform a linear input signal into a non-linear output signal. It is the most commonly used activation function and is defined as f(x)=max(0, x). ReLU is simple to compute, as it requires only one computation step. It is also computationally efficient and has been found to improve the performance of deep learning networks.

6 RESULTS AND DISCUSSION

6.1 EVALUATION METRICS

Evaluation metrics are used to measure the effectiveness and performance of predictive models. A predictive model is often trained using a particular algorithm or approach on a set of data, referred to

 DOI: 10.36893.JK.2023.V13I04N16.0033-0038 Copyright @ 2023 Author

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

as the training dataset. When the model has been trained, it is tested using a different dataset known as the holdout dataset that wasn't utilised for training.

The model makes predictions on the holdout dataset during the testing phase, and the predictions are contrasted with the holdout dataset's expected values. The model's performance is then measured using evaluation criteria depending on how precisely it makes predictions.

In general, evaluation metrics are crucial for assessing the efficacy of predictive models and assisting in providing direction for model performance enhancements.

6.2 ANALYSIS

	75%-25%	80%-20%	90%-10%
Precision	0.94	0.97	0.95
Recall	0.93	0.94	0.93
Fi-score	0.93	0.95	0.94

Fig 5. OUTPUT IF NO DATA IS GIVEN

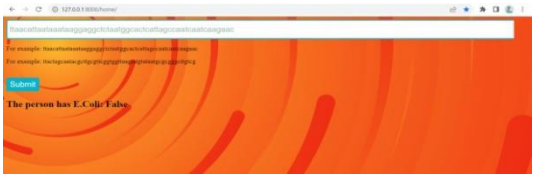

Fig 6. OUTPUT FOR NO E. COLI

Fig 7. OUTPUT FOR E. COLI

6.2 EXPERIMENTAL SETUP

Device name: DESKTOP1234
Processor:12th Gen Intel(R) Core (TM) i5-1235U 1.30 GHz
Installed RAM:8.00 GB (7.68 GB usable)
Device ID: CEB56CE5-54A5-44F9-85C2-F88EEC22C6EA
Product ID:00356-24570-39239-AAOEM
System type:64-bit operating system, x64-based processor

6.3 WINDOWS SPECIFICATION

Edition: Windows 11 Home Single Language
Version:22H2
Installed on:08-10-2022
OS build:22621.1413
Experience: Windows Feature Experience Pack 1000.22639.1000.0

DOI: 10.36893.JK.2023.V13I04N16.0033-0038
Copyright @ 2023 Author

ISSN: 2278-4632
Vol-13, Issue-04, No.06, April : 2023

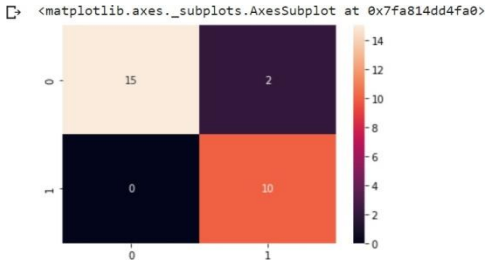

Fig 8. CONFUSION MATRIX

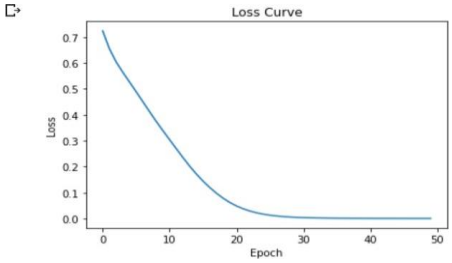

Fig 9. LOSS CURVE

10 CONCLUSIONS

In conclusion, the proposed system offers an efficient and accurate way of identifying the presence of E. coli bacteria in a DNA sample. The MLP classifier model that has been developed is accurate in classifying the DNA sample into E. coli bacteria or not. Furthermore, the website was developed with user-friendly features that allow users to input their DNA samples and get the results of the classification in real-time. This system is useful for medical purposes, as it can provide quick and accurate results. In addition, it is also useful for educational purposes, allowing students to learn about the technique of virus identification in a DNA sample.The future of this project could involve expanding its capabilities to identify not only E. coli bacteria but other related bacteria as well. This would allow the system to be more widely used by medical professionals and researchers for the identification and classification of various microorganisms. Furthermore, the system could be further improved by exploring different machine learning and deep learning models to ensure better accuracy and faster classification results. Finally, the system could be adapted to be used in different filed, such as the detection of genetic diseases and the identification of food-borne bacteria.

REFERENCES

Hensel, M., J. E. Shea, C. Gleeson, M. D. Jones, E. Galton, and D. W. Holden.1995. simultaneous identification of bacterial virulence genes by negative selection. Science 269:400-403.

Handfield, M., L. J. Brady, A. Progulse-Fox, and J.D. Hillman. 2000. IVIAT: a novel method to identify microbial genes expressed specifically during human infection. Trends Microbiol. 8:336-339.

Berg, R., & Doolittle, R. F. (1982). A hierarchical classification of bacterial viruses. Journal of Molecular Evolution, 18(3), 224-229.

Chaudhuri, S., & Chaudhuri, S. (2015). Machine Learning for microbial genomics. Trends in Genetics, 31(9), 481-490.

Fiehn, O. (2002). Metabolomics- the link between genotypes and phenotypes. Plant Molecular Biology,48(1-2), 155-171.

LeCun, Y., Bengio, y., &Hinton, G. (2015). Deep Learning. Nature, 521(7553), 436-444.

UCI Machine Learning Repository (2020). E. coli promoter data set. Retrieved from: https://archive.ics.uci.edu/ml/machine-learning-datasets/molecular-biology/promoter2-gene-sequences/

Wirawan, D. (2018). Machine learning for medical diagnosis and treatment. International Journal of Computer Science & Applications, 15(3), 4-1.

https://www,ncbi.nlm.nih.gov/pmc/articles/PMC128117/

https://www.sciencedirect .com/science/article/abs/pii/S0168160500002063? via%3Dihub

 DOI: 10.36893.JK.2023.V13I04N16.0033-0038 Copyright @ 2023 Author