iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45866

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Forecast of mRNA Articulation in Cows Milk Utilizing mRNA Auxiliary Designs and ML Classifiers

Mrs.Jyothi G.C[1], Kruthika B M[2], Sushma K M[3], Apoorva B R[4], Vijayalaxmi[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*U.G.Scholor, Information Science and Engineering Bapuji Institute of Engineering and Technology, Davangere,India*

Abstract: ***The mRNA molecules expressed in cow's milk are important molecular biomarkers for different physiological and pathological conditions in cattle. The prediction of the quantity that a specific mRNA type could be expressed in cow's milk is a challenging theoretical task. The current study presents for the first time several different Machine Learning models to predict the mRNA expression using the mRNA secondary structure fragments.***
Keywords: mRNA expressions, mRNA secondary structure fragments, Machine Learning, LSTM and GRU Models.

I. INTRODUCTION

The mRNA expression in cow's milk is an important biomarker for the cattle conditions. The current study proposes a method to predict the low or high expression levels of mRNA using mRNA secondary structure fragments and Machine Learning classifiers. Essentially, the terms "classifier" and "model" are synonymous in certain contexts; however, sometimes people refer to "classifier" as the learning algorithm that learns the model from the training data. Model: In machine learning field, the terms hypothesis and model are often used interchangeably. In other sciences, they can have different meanings, i.e., the hypothesis would be the "educated guess" by the scientist, and the model would be the manifestation of this guess that can be used to test the hypothesis. Classifier: A classifier is a special case of a hypothesis (nowadays, often learned by a machine learning algorithm). A classifier is a hypothesis or discrete-valued function that is used to assign (categorical) class labels to particular data points. In the email classification example, this classifier could be a hypothesis for labeling emails as spam or non-spam. However, a hypothesis must not necessarily be synonymous to a classifier.

II. *PROBLEM* STATEMENT

The risk of infectious state of the DNA might lead to less accuracy in prediction the quality of the milk. Efficiency is less with the usage of different therapeutic tools for predicting. The cost for predicting, using different protein is more compared to using mRNA tool.

III. PROPOSED SYSTEM

The proposed system is to predict the secondary structure of mRNA in cow's milk using final dataset of 30 selected features which becomes the input of the machine learning technique called as "Recurrent neural network" in Google Colab.

IV. OBJECTIVES

1) To collect the data set from the standard database.
2) Preprocessing of data to improve upon the quality of dataset.
3) By using machine learning classifier LSTM/GRU can predict low/high expression of new mRNA types in cow's milk.

V. LITERATURE SURVEY

COVID-19 mRNA Vaccine Degradation Prediction Using LR and LGBM Algorithms Soon Hwai Ing, Azian Azamimi Abdullah, Nor Hazlyna Harun and Shigehiko Kanaya proposed a paper of COVID-19 mRNA Vaccine Degradation Prediction Using LR and LGBM Algorithms The threatening Coronavirus which was assigned as the global pandemic concussed not only the public health but society, economy and every walk of life. Some measurements are taken to stifle the spread and one of the best ways is to carry out some precautions to prevent the contagion of SARS-cov-2 virus to uninfected populaces.

Injecting prevention vaccines is one of the precaution steps under the grandiose blueprint. Among all vaccines, it is found that mRNA vaccine which shows no side effect with marvellous effectiveness is the most preferable candidates to be considered. However, degradation had become its biggest drawback to be implemented. Hereby, this study is held with desideratum to develop prediction models specifically to predict the degradation rate of mRNA vaccine for COVID-19. Two machine learning algorithms, which are, Linear Regression (LR) and Light Gradient Boosting Machine (LGBM) are proposed for models development using Python language. Dataset comprises of thousands of RNA molecules that holds degradation rates at each position from Eterna platform is extracted, pre-processed and encoded with label encoding before loaded into algorithms. The results show that LGBM (0.2447) performs better than LR (0.3957) for this study when evaluate with the RMSE metric.

Prediction of mRNA expression in cow's milk using mRNA secondary structures and Machine Learning classifiers Rodrigo Martín, Yong Liu, Omar Landaeta , Luis Felipe Llamas , Chuanshe Zhou, Zhiliang Tan, Haibo Zhang , Cristian R Munteanu, proposed a paper of Prediction of mRNA expression in cow's milk using mRNA secondary structures and Machine Learning classifiers the prediction of the quantity that a specific mRNA type could be expressed in cow's milk is a challenging theoretical task. The current study presents for the first time several different Machine Learning models to predict the mRNA expression using the mRNA secondary structure fragments. This unique methodology is based on a dataset of experimental mRNA expression data. Thus, the best classification model was obtained with bayes net method and is based on 24features and 4067 cases. The model has the true positive rate for the low mRNA expression class of 0.78 (average true positive rate of 0.66). Further studies are needed improve the current results, using datasets with different feature sets and more advanced Machine Learning methods.

Prediction of mRNA subcellular localization using deep recurrent neural networks: Zichao Yan1, Eric Le´ cuyer and Mathieu Blanchette, Prediction of m subcellular localization using deep recurrent neural networks: Messenger RNA proposed a paper of subcellular localization mechanisms play a crucial role in posttranscriptional gene regulation. This trafficking is mediated by trans-acting RNA-binding proteins interacting with cis¬ regulatory elements called zip codes. While new sequencing-based technologies allow the high¬ through put identification of RNA localized to specific subcellular compartments, the precise mechanisms at play, and their dependency on specific sequence elements, remain poorly understood introduce RNA tracker, a novel deep neural network built to predict, from their sequence alone, the distributions of mRNA transcripts over a predefined set of subcellular compartments. RNA tracker integrates several states of the art deep learning techniques (e.g., CNN, LSTM and attention layers) and can make use of both sequence and secondary structure information. We report on a variety of evaluations showing RNA tracker's strong predictive power, which is significantly superior to a variety of baseline predictors. Despite its complexity, several aspects of the model can be isolated to yield valuable, testable mechanistic hypotheses, and to locate candidate zip code sequences within transcripts.

VI. SYSTEM DESIGN

System design thought as the application of theory of the systems for the development of the project. System design defines the architecture, data flow, use case, class, sequence and activity diagrams of the project development. This architecture diagram illustrates how the system is built and is the basic construction of the software method. Creations of such structures and documentation of these structures is the main responsible of software architecture.

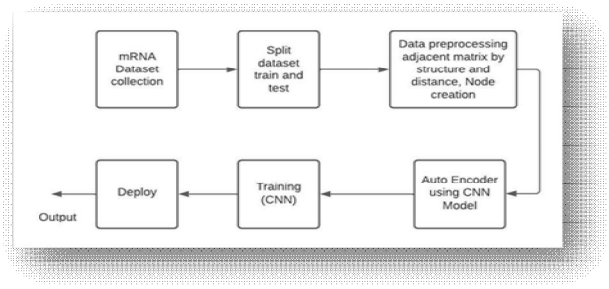


Fig: System Architecture

A. Collection of Dataset

Dataset was collected by Kaggle website. The sample of mRNA secondary structure sequences are undergone high throughput Screening process. It is having features like reactivity, deg-pH - mg, deg- mg, 50c and id of the mRNA secondary structure.Submission.csv is the original dataset. From this data file we split the dataset into train and test dataset. Bpps folder is having the mRNA structure Id in .npy file format

1) *Bpps folder:* This folder consists id of the mRNA Secondary Structure. The id files of the mRNA Secondary Structure were saved in .npy format. • Submission File: This is the original dataset with features like reactivity, deg_Mg_pH10, deg_pH10, deg_Mg_50C, deg_50C and id_seqpos of the mRNA structure. The file saved in .csv format.
2) *Train dataset:* The train dataset consists 2026 data with 19 features and saved in .csv. The features present in train dataset are as follows:
id_seqpos,sequence,predicted_loop_type,Signal_to_noise,SN_filter,Seq_Scored,Reactivity_error,deg_error_Mg_pH10,deg_error_pH10,deg_Mg_error_50C,de g_error_50C,reactivity,deg_Mg_pH10,deg_pH10,deg_Mg_50C, deg_50C.
3) *Test Dataset:* The test dataset consists 3634 data with 6 features and saved in .csv. The features present in train dataset are reactivity, sequence, structure, predicted_loop_type, seq_length, seq_scored.
4) *Dataset Features*

a) deg Mg pH10 - (1×68 vector in Train and Public Test, 1x91 in Private Test) An array of floating-point numbers, should have the same length as seq_scored. These numbers are reactivity values for the first 68 bases as denoted in sequence, and used to determine the likelihood of degradation at the base/linkage after incubating with magnesium in high pH (pH 10).
b) deg 50C - (1x68 vector in Train and Public Test, 1x91 in Private Test) An array of floating-point numbers, should have the same length as seq_scored. These numbers are reactivity values for the first 68 bases as denoted in sequence, and used to determine the likelihood of degradation at the base/linkage after incubating without magnesium at high temperature (50 degree Celsius).
c) deg Mg_50C - (1x68 vector in Train and Public Test, 1x91 in Private Test) An array of floating-point numbers, should have the same length as seq_scored. These numbers are reactivity values for the first 68 bases as denoted in sequence, and used to determine the likelihood of degradation at the base/linkage after incubating with magnesium at high temperature (50 degrees Celsius).
d) error - An array of floating-point numbers, should have the same length as the corresponding reactivity or deg_* columns, calculated errors in experimental values obtained in reactivity and deg_* columns.
e) predicted_loop_type - (1×107 string) Describes the structural context (also referred to as 'loop type') of each character in sequence. Loop types assigned by bp RNA from Vienna RNA fold 2 structure. From the bp RNA documentation: S: paired "Stem" M: Multiloop I: Internal loop B: Bulge H: Hairpin loop E: dangling End X: external loop vi.
f) id -An arbitrary identifier for each sample.
g) seq_scored - (68 in Train and Public Test, 91 in Private Test) Integer value denoting the number of positions used in scoring with predicted values. This should match the length of reactivity, deg_ and error* columns. Note that molecules used for the Private Test will be longer than those in the Train and Public Test data, so the size of this vector will be different.
h) seq_length - (107 in Train and Public Test, 130 in Private Test) Integer values, denotes the length of sequence. Note that molecules used for the Private Test will be longer than those in the Train and Public Test data, so the size of this vector will be different.
i) sequence - (1 * 107 string in Train and Public Test, 130 in Private Test) Describes the RNA sequence, a combination of A, G, U, and C for each sample. Should be 107 characters long, and the first 68 bases should correspond to the 68 positions specified in seq_scored (note: indexed starting at 0).
j) structure - (1 * 107 string in Train and Public Test, 130 in Private Test) An array of (,) and characters that describe whether a base is estimated to be paired or unpaired. Paired bases are denoted by opening and closing parentheses e.g. (....) means that base 0 is paired to base 5, and bases 1-4 are unpaired.
k) reactivity - (1 * 68 vector in Train and Public Test, 1 * 91 in Private Test) An array of floating-point numbers, should have the same length as seq_scored. These numbers are reactivity values for the first 68 bases as denoted in sequence, and used to determine the likely secondary structure of the RNA sample.
l) deg pH10 (1 * 68 vector in Train and Public Test, 1 * 91 in Private Test) An array of floating-point numbers, should have the same length as seq_scored.

B. Data Preprocessing

1) By using the SN filter remove the noise present in the dataset and plot the graph of signal to noise distribution and SN filter distribution graph.
2) After filtering techniques calculate the summation of all the Id of mRNA, calculate the maximum value of the dataset and the mean of the mRNA sequence.

3) The augmentation technique is for the extraction of mRNA features on the whole projection hence 31 features were get extracted of the mRNA prediction take place.
4) After augmentations the dataset will changes. Make the sequence batch wise for the training purpose by using tokenization technique. Here the length of the tokenization is 14.

C. GRU

1) The gated recurrent neural network is one of the RNN model to predict the output. It takes input the current input and the previous hidden state as vector.
2) GRU or Gated recurrent unit is an advancement of the standard RNN i,e recurrent neural network.
3) Manipulating the sequence and extract the features of mRNA and build the GRU model. Here there is memory cell for the storing and inviting the information. It shows the architecture of GRU unit.

D. LSTM

1) The long short-term memory is one of the RNN model for the prediction of mRNA sequence.
2) It is one the model to solve the complex problems in machine learning. It is having the memory cell to modify the mRNA sequence based on the information stored in the memory cell.
3) After training process, the models are developed. By using the training history predict the mRNA test sequence and the output.

VII. DATA FLOW DIAGRAM

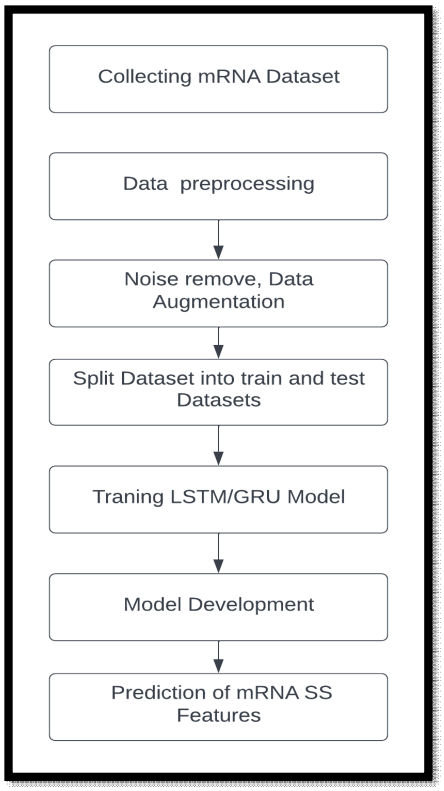


Fig: Flow Diagram

A. Collecting of mRNA Dataset:

Dataset was collected by Kaggle website. The sample of mRNA secondary structure sequences are undergone high throughput Screening process. It is having features like reactivity, deg-pH - mg, deg- mg, 50c and id of the mRNA secondary structure.Submission.csv is the original dataset. From this data file we split the dataset into train and test dataset. Bpps folder is having the mRNA structure Id in .npy file format.

B. Data Preprocessing

After getting the data we start with data pre-processing. Data preprocessing includes:

1) Checking whether the features have missing values or not in train test and submission dataset.

2) By using the SN filter remove the noise present in the dataset and plot the graph of signal to noise distribution and SN filter distribution graph. We obtained samples with signal to noise greater than 1 is of 2096. Samples with SN_Filter equals to 1 is of 1589 and samples with signal to noise greater than1, but SN_Filter equals to 0 is 509.

3) After filtering techniques calculate the summation of all the Id of mRNA , calculate the maximum value of the dataset and the mean of the mRNA sequence.

4) The data augmentation technique is applied on test and train dataset. Before applying the data augmentation samples in train dataset is 2400 and in samples in test dataset is 3634. After applying the data augmentation 20 samples are increased in train dataset and 20 data samples are increased in test dataset. The augmentation technique is for the extraction of mRNA features on the whole projection hence 31 features were get extracted of the mRNA prediction take place.

5) In data augmentation we got some features on base of Id, sequence, structure of mRNA like log_gamma, score, cnt.

6) Make the sequence batch wise for the training purpose by using tokenization technique. Here the length of the tokenization is 14.

C. Training LSTM/GRU Model

1) We applied the LSTM model for training the datasets by specifying with 50 Epochs after training the model with train datasets we got the 0.1438 losses.

2) We applied the GRU model for training the datasets by specifying with 50 Epochs after training the model with train datasets we got the 0.1463 losses.

3) Model development: We develop the model by training with the samples of train dataset and tested with samples of test dataset for evaluating the accuracy in both GRU and LSTM model. GRU and LSTM model development, predict the mRNA secondary structure sequence and the display the output.

4) Predictions: The test dataset is applied to LSTM / GRU Model. The feature of development model compares with the test dataset features and extract the secondary structure information of mRNA.

VIII. IMPLEMENTATION

Algorithms used are LSTM and GRU

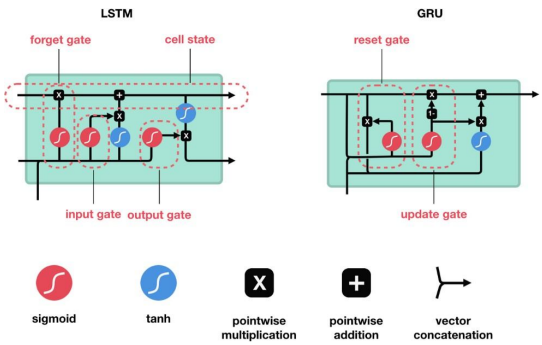


Fig: LSTM and GRU Model

A. LSTM

An LSTM has a similar control flow as a recurrent neural network. It processes data passing on information as it propagates forward. The differences are the operations within the LSTM's cells.

1) Sigmoid: Gates contains sigmoid activations. A sigmoid activation is similar to the tanh activation. Instead of squishing values between -1 and 1, it squishes values between 0 and 1. That is helpful to update or forget data because any number getting multiplied by 0 is 0, causing values to disappears or be "forgotten."

2) Tanh activation: The tanh activation is used to help regulate the values flowing through the network. The tanh function squishes values to always be between -1 and 1.

3) Forget gate: This gate decides what information should be thrown away or kept. Information from the previous hidden state and information from the current input is passed through the sigmoid function. Values come out between 0 and 1. The closer to 0 means to forget, and the closer to 1 means to keep.

4) *Input Gate:* First, we pass the previous hidden state and current input into a sigmoid function. That decides which values will be updated by transforming the values to be between 0 and 1. 0 means not important, and 1 means important. You also pass the hidden state and current input into the tanh function to squish values between -1 and 1 to help regulate the network. Then you multiply the tanh output with the sigmoid output. The sigmoid output will decide which information is important to keep from the tanh output.
5) *Cell State:* The cell state gets pointwise multiplied by the forget vector. This has a possibility of dropping values in the cell state if it gets multiplied by values near 0. Then we take the output from the input gate and do a pointwise addition which updates the cell state to new values that the neural network finds relevant. That gives us our new cell state.
6) *Output Gate:* The output gate decides what the next hidden state should be. Remember that the hidden state contains information on previous inputs. The hidden state is also used for predictions. First, we pass the previous hidden state and the current input into a sigmoid function. Then we pass the newly modified cell state to the tanh function. We multiply the tanh output with the sigmoid output to decide what information the hidden state should carry. The output is the hidden state. The new cell state and the new hidden is then carried over to the next time step. GRU The GRU is the newer generation of Recurrent Neural networks and is pretty similar to an LSTM. GRU's got rid of the cell state and used the hidden state to transfer information. It also only has two gates, a reset gate and update gate.
7) *Update Gate:* The update gate acts similar to the forget and input gate of an LSTM. It decides what information to throw away and what new information to add.
8) *Reset Gate:* The reset gate is another gate is used to decide how much past information to forget.

B. Algorithm Applied

1) *Step 1:* Collecting all the libraries like Keras, Pandas, NumPy, Sklearn.
2) *Step 2:* Collection of datasets of mRNA sequence.
3) *Step 3:* Define the target value.
4) *Step 4:* Data preprocessing.
a) Data augmentation.
b) SN filter noise remove.
c) Tokenization.
5) *Step 5:* Apply dataset to LSTM and GRU model.
6) *Step 6:* Model build.
7) *Step 7:* Prediction.
8)

IX. EXPERIMENTAL RESULTS

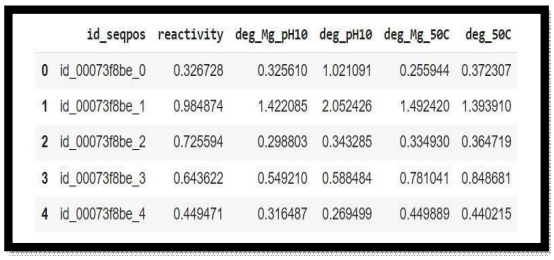
A. Outputs

After the GRU and LSTM model development, predict the mRNA secondary structure sequence and the display the output. Both LSTM and GRU model are used to predict the sequence and the accuracy's 86%.

B. Results Snapshots

	id_seqpos	reactivity	deg_Mg_pH10	deg_pH10	deg_Mg_50C	deg_50C
0	id_00073f8be_0	0.369184	0.345683	1.042123	0.280313	0.411077
1	id_00073f8be_1	1.170666	1.609773	2.282659	1.652078	1.506991
2	id_00073f8be_2	0.772977	0.303126	0.337532	0.325237	0.355706
3	id_00073f8be_3	0.688509	0.588388	0.626373	0.834521	0.910515
4	id_00073f8be_4	0.455789	0.307601	0.290361	0.444387	0.455857

Fig: Predictions of GRU



	id_seqpos	reactivity	deg_Mg_pH10	deg_pH10	deg_Mg_50C	deg_50C
0	id_00073f8be_0	0.326728	0.325610	1.021091	0.255944	0.372307
1	id_00073f8be_1	0.984874	1.422085	2.052426	1.492420	1.393910
2	id_00073f8be_2	0.725594	0.298803	0.343285	0.334930	0.364719
3	id_00073f8be_3	0.643622	0.549210	0.588484	0.781041	0.848681
4	id_00073f8be_4	0.449471	0.316487	0.269499	0.449889	0.440215

Fig: Predictions of LSTM

X. CONCLUSION

This project is for the predictions of mRNA secondary structure using the mRNA secondary structure dataset. The prediction of mRNA structure secondary sequence information done by using GRU and LSTM module where these modules are undergone training the dataset and predict the correct mRNA secondary structure sequences for testing data. Both the system predicted same accuracy and can prefer one of the models

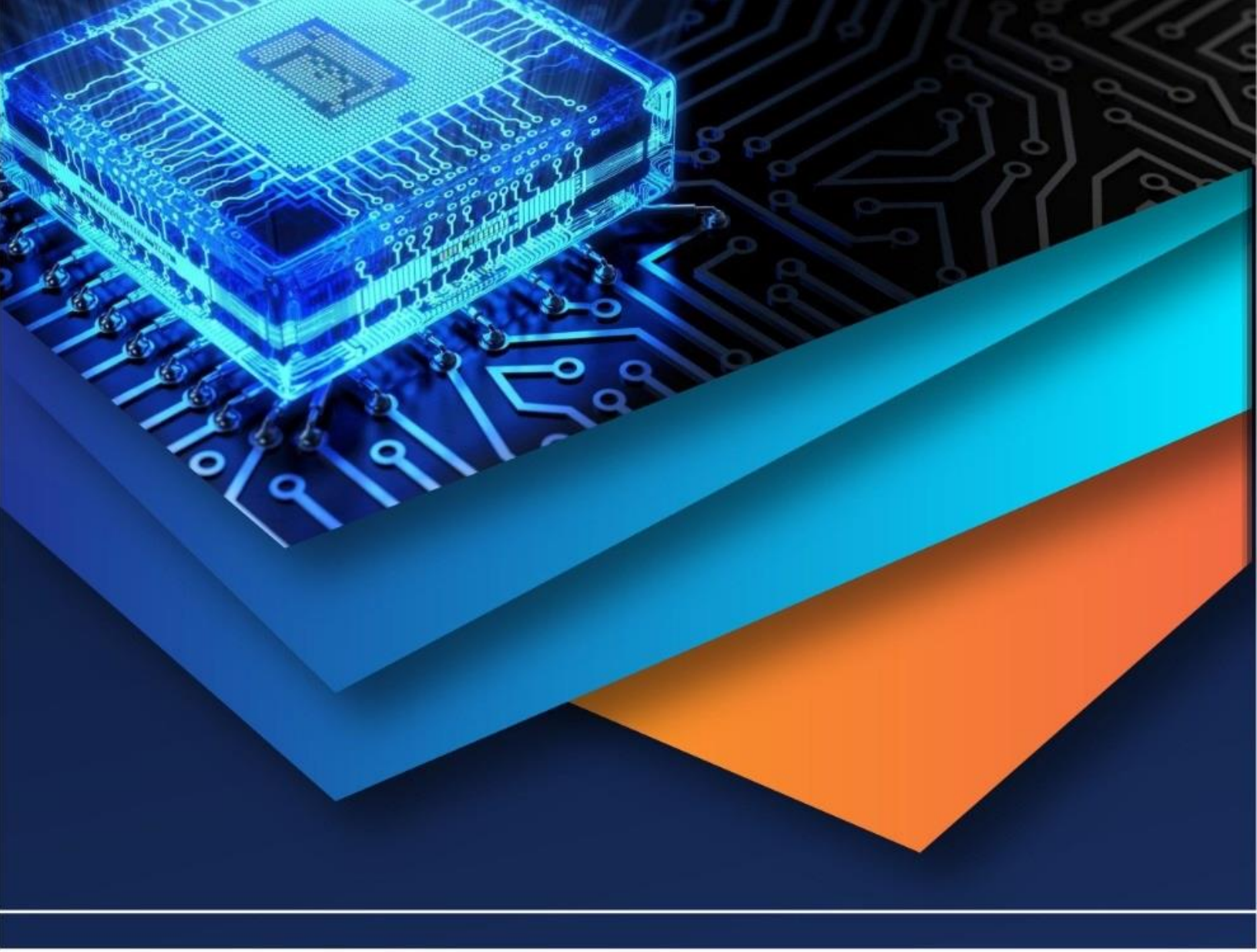
REFERENCE

[1] RE1.Murrieta, C.M.; Hess, B.W.; Scholljegerdes, E.J.; Engle, T.E.; Hossner, K.L.; Moss,G.E.;Rule, D.C. Evaluation of milk somatic cells as a source of mRNA for study of lipogenesisin the mammary gland of lactating beef cows supplemented with dietary high-linoleate safflower seeds.J. Anim. Sci. 2006, 84,2399-2405.

[2] Ma, J.L.; Zhu, Y.H.; Zhang, L.; Zhuge, Z.Y.; Liu, P.Q.; Yan, X.D.; Gao, H.S.; Wang, J.F.Serum concentration and mRNA expression in milk somatic cells of toll- like receptor 2, toll-like receptor 4, and cytokines in dairy cows following intramammary inoculation witheschericha coli.J. Dairy Sci. 2011, 94,5903-5912.

[3] Witten, I.; Frank, E. Data mining: Practical machine learning tools and techniques, second edition (morgan kaufmann series in data management systems). Morgan Kaufmann: 2005.

[4] Smith, T.C.; Frank, E. Introducing machine learning concepts with weka. In Statistical genomics: Methods and protocols, Springer: New York, NY, 2016; pp 353-378



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY