IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 08, 2022

A COMPARATIVE STUDY OF FAKE NEWS DETECTION BETWEEN MACHINE LEARNING AND DEEP LEARNING APPROACHES

Vennam Ratna Kumari, Dr.Mula Veera

Computer science and engineering, Nigama Engineering College Karimnagar Telangana

Hanumantha reddy, Dept of Computer science and engineering, Vivekananda institute of technology and sciences

ratnab@nigama.org, Hanuman.vits@gmail.com

Abstract: The majority of people these days prefer to read the news online via social media. A multitude of websites disseminate news and offer the source of verification. How to validate news and articles shared on social media platforms like Facebook Pages, Twitter, WhatsApp groups, and other microblogs and social networking sites is the question. It is detrimental to society for people to trust rumors and pass them off as news. It's vital to put an end to the rumors and concentrate on accurate, verified news reports. This project's goal is to create two models that use machine learning (SVM) and deep learning (LST) algorithms, respectively, to identify bogus news.

Utilizing both machine learning and deep learning, an attempt is made to aggregate news and then use Support Vector Machine and Long Short-Term model to determine whether the news is real or fake. First, the dataset is cleaned and pre-processed; next, feature extraction techniques are applied to the pre-processed data, and the model is trained using both algorithms independently to obtain two distinct models using SVM and LSTM, respectively. Confusion Matrix, Classification reports, and accuracies of both models are calculated and compared in order to identify the best model for Fake News Detection.

The accuracy of the LSTM classifier was 99.54%, while the SVM classifier yielded 99.29%. While both algorithms produced results with acceptable accuracy, LSTM performed better in classifying the news articles than SVM. This project's main contribution is comparing the accuracies of SVM and LSTM algorithms to determine which algorithm and techniques match the problem of fake news detection the best.

Keywords: SVM, LSTM, Machine Learning, Deep Learning, Fake News.

I. INTRODUCTION

People use social media nowadays for business, education, pleasure, and informational purposes. With so much news and information available, the question of whether it is real or fake always arose. The dissemination of fake news frequently has the intention of misleading or inciting a desire for financial or political gain.

A survey conducted in July 2020 revealed that 59% of people worldwide were internet users. The goal of fake news is to distort readers' perceptions. As an obvious example, consider the COVID-19 epidemic, which is rapidly spreading over the world at this time and giving rise to false information about the illness in our society. Such false information led to people's uneasiness and even had fatal health implications.

It is becoming more and more difficult to distinguish the truth from the fake due to the internet's exponential expansion in information. This brings up the issue of false information. The proliferation of technology in today's world has led to a quick production of new news content, making the detection of fake news online crucial. Therefore, finding fake news is a necessary task. Shows what is real and what is fake news.

Examples of Spread of Fake News:

While some false news is harmless, others have the potential to do serious harm because they promote anti-democratic ideas. Experts believe that social bots and fake news have had a significant impact on global events. Here are a few instances of "successful" fake news that impacted opinion around the world:

- During the 2020 election in our nation, a large amount of false material spread on social media platforms with the intention of distorting public opinion as well as that of organizations and political parties.
- **AIDS conspiracy:** False information was disseminated through blind media faith even before information became digital. In the 1980s, West German media disseminated the story that the US secret service CIA had produced the AIDS virus on behalf of secret services run by the GDR and the Soviet Union. We refer to this as a disinformation campaign.
- **The Bitcoin scam:** Since cryptocurrencies have become more and more popular, there have been more frauds and scams. Claimed bitcoin trading platforms used made-up testimonials from well-known individuals to promote their services and gain public trust. The reviews advised prospective investors to expect large returns.
- **Fake news about the immigration crisis:** A lot of false information was disseminated to divide the European population during the immigration crisis. For instance, a report in the British daily "Daily Express" in February 2017 stated that Germany intended to import 12 million migrants. Correctiv.org, a German investigative newsroom, later proved this to be false.

Today, states, organizations, and individuals have used fake news publicity on the internet for a variety of purposes. Social media is frequently used to create and disseminate spectacular news in order to accomplish desired results. However, it could also entail an intentionally exaggerated narrative of a factual event. To catch readers' attention, this may also entail naming the webpages with deceptive titles or taglines. Such false information could result in criminal activity, societal unrest, financial frauds due to deception, political advantage, an increase in readers, click-through revenue, etc. This could have an impact on the significance of credible news sources as well. Another risk is that other electronic media may use this as a source for news, which would encourage the news to spread even further. Determining the veracity of news and internet content is the issue. The difficulty of identifying the bots that distribute misleading information is equally significant. The objective of this effort is to develop a model that can forecast the likelihood of a news report being fabricated or not based on historical data. We create two models, one for machine learning (using SVM) and the other for deep learning (using LSTM), to identify bogus news. The objective of our experiment is to determine which of the two

built models the SVM and LSTM approaches is more accurate in classifying fake news. By determining whether the news is true or not, this project determines whether the news story is from a reliable source. To detect fake news, we suggest using deep learning models and machine learning approach models. The best fit model for categorizing news as true or fake is determined by comparing the accuracy of two models.

II. LITERATURE SURVEY

Sobhani, Parinaz, Saif Mohammad, and Svetlana Kiritchenko et.al. [1] released a study on the use of SVM ngrams for fake news detection. In this work, SVM ngram features were utilized to identify bogus news in tweets related to SemEval 2016 data. An N-gram is a length-n continuous series of elements. Employing N-gram results in more restrictive classifiers. In comparison to the taste attitude, this SVM model produced a 10% lower score for the anti stance, indicating a small imbalance in categorization.

Davis, Richard, and Chris Proctor et.al. [2] created a research paper for BoW MLP-Based Fake News Detection. He claimed in this that the Bag of Word Multilayer Perceptron was used in the development of the model for the detection of bogus news. Prior to classifying, the corpus is first transformed into a bag of words and vectors. Two SoftMax layers—one for relevance and the other for attitudes—as well as an entropy-based cost function are then used. Using BoW, this model has an 80% accuracy rate. Using TF-IDF can help with this.

Alim Al Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, Myung Suh Choi et.al.[3] suggested a research piece titled "Online fake news detection using KNN," in which it was explained how machine learning classifiers, which may assist in determining whether the news is true or fraudulent, can be used to quickly distinguish and detect fake news. In this model, K-Nearest Neighbor and SVM classifiers are utilized. They have a 74% accuracy rate thanks to their classifiers.The application of unsupervised machine learning classifiers for the identification of fake news may be the subject of future research. They have a 74% accuracy rate thanks to their classifiers.The application of unsupervised machine learning classifiers for the identification of fake news may be the subject of future research.

Rubin ,K. Shu, A.Sliva, S.Wang, J. Tang and H. Liu et.al.[4] Using Naïve Bayes, SVM, and neural networks, they studied how news might be categorized as true or false by concentrating on a few characteristics that are frequently found in fake news. These attributes, according to them, were predicated on "representative datasets, evaluation metrics, existing algorithms from a data mining perspective, and psychology and social theories." They added that the current model only functions with the dataset that already exists; it is not appli"able to newly collected data. In order to further enhance the model, the next stage would be to train it and examine how the accuracies change with the addition of fresh data.

Krejzl and Steinberger et.al.[5] suggested a literature review on the use of several domain knowledge factors in the detection of bogus news online. The usage of domain knowledge-related characteristics in SemEval 2015 data, a tweet-based stance identification job, was the main emphasis of this research. To extract syntactic and

semantic features, part-of-speech tags, general inquiry, and entity-centered sentiment dictionaries were used. Additionally, they developed a domain stance dictionary that enumerates the terms that appear most frequently in each stance. They also reported that, while their model produced good accuracy scores, it fell short of the SVM ngram model.

III. SYSTEM ANALYSIS

Two models are created in our proposed system to detect fake news using Machine Learning algorithms (SVM) and Deep Learning algorithms (LSTM), respectively. The suggested system is provided the news that we wish to categorize as true or fraudulent. Next, the suggested systems categorize the provided news as either Real or Fake news. The dataset consists of two CSV files, one containing true news and the other fraudulent news. The data is pre-processed after it has been cleansed, that is, missing values and noisy data removed. By vectorizing the pre-processed data using TF-IDF methods, the features are extracted. This gives the data's significance or synopsis. The SVM algorithm, an ML technique, is used to train the model, while the LSTM algorithm, a DL approach, is used to train the other model. The vectorized data is processed by these two models, which then identify whether the news is real or fake.

IV. INTERFACE/SOFTWARE REQUIREMENTS

Equipment Interfaces: The program is designed to function as a stand-alone, single-client system. A laptop will be used to execute the application. No additional devices or connection points will be needed.

1. Programming Interfaces

2. The product will receive input from a single source. First things first: the UI. The investigation meeting and the text will be provided by the UI.

3. Produces text as the output format.

4. Working System.

5. UIs In order to accommodate the needs of the clients, the point of engagement will fulfill the ancillary requirements. It will be simple and uncomplicated. Controls that allow the user to interact with the program will be obvious and suggest their purpose inside the program. The connection point will include two illustrations that are arranged underneath in addition to client inputs.

V. METHODOLOGY

OVERVIEW OF A DATA SET

Two CSV file datasets from Kaggle—one for actual news and the other for fake news—make up the data used to train the models. The four attributes in the dataset provide information about the news stories' title, text, subject, and date. True CSV files include real news data, while false CSV files contain fake news data.

Type of Articles	Total Count	Total Attributes
Real	21417	4
Fake	23481	4

Table: Dataset's Quantitative Analysis

DATA CLEANING AND PRE-PROCESSING

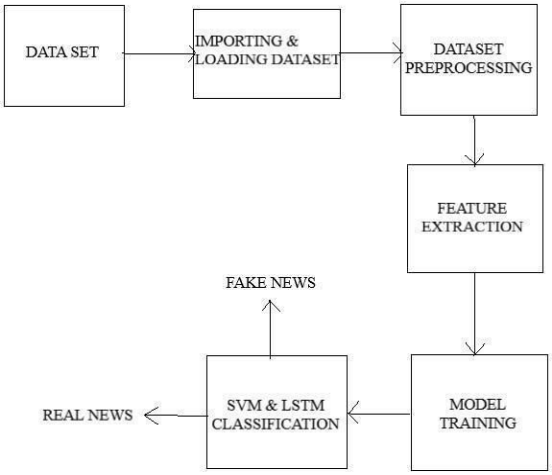
Eliminating noisy data from a file that contains null or missing values is known as data cleaning. Prior to pre-processing the data, it must be completed. In order to ensure that the data used to develop

machine learning models is consistent and reliable, pre-processed data is an essential stage in the data science lifecycle. News on social media is large, loud, and unstructured. Pre-processing of these data is therefore necessary. First, we removed duplicate rows and missing values from the dataset during the pre-processing stage of the data. Following that, we eliminated stop words, digits, alphanumeric text, punctuation, and non-English terms from the news stories. These values don't actually add any value to the news content; instead, they may lead to the model overfitting the data.

Feature Extraction: Only numerical values can be handled by machine learning models. As a result, we must change the text input data into a vector form of numbers so that classification techniques can be used. Vectorization is the process of transforming textual material into a numerical representation. In text processing, there are numerous word vectorization algorithms. Our model makes use of pre-trained word embedding techniques and the frequency-based method (TF-IDF) among them.

TF-IDF: This sparse vector representation assesses a word's level of relevance within a corpus, or group of documents. All of the sentence's tokens are used as vocabulary in TF-IDF approaches. phrase frequency, or TF, is a metric that indicates how frequently a phrase appears in a single text. IDF calculates a term's importance within a document.

SYSTEM ARCHITECHTURE



WORD EMBEDDING

Words with a similar significance are addressed much the same way in a message thanks to a learned portrayal called word implanting. It helps with changing over text record characters — crude information — into a significant arrangement of word vectors in the implanting space so the model can work all the more proficiently. One of the significant advances in profound learning for troublesome regular language handling issues might be this strategy for communicating words and reports.

VI. MODEL TRAINING

SVM MODEL:

A calculation for managed learning is a help vector machine (SVM). SVMs capability by being prepared on specific information that has previously been separated into two gatherings. Subsequently, the model is assembled whenever it has been prepared. The SVM approach should likewise amplify the edge between the two classes as well as figuring out which classification any new information has a place with. The SVM's capacity to distinguish a hyperplane that parts the dataset into two gatherings is great.

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 08, 2022

A piece capability is a method for taking information and changing it into the organization required for information handling. At the point when the information is directly detachable — that is, the point at which it very well may be separated utilizing a solitary line — the program utilizes a straight part. It is among the most broadly used pieces. It is normally applied when a given Informational index contains countless highlights.

LSTM MODEL

This particular kind of recurrent neural network processes both time series and sequential input. LSTM is better suited for text categorization than machine learning methods since it consists of three gates and can capture long-term connections between word sequences. A cell state limits the number of linear interactions permitted, allowing the data to pass through the cell units unaltered. In an LSTM, the output gate is used to calculate the activation output of the LSTM cell, the forget gate regulates the amount of time a value stays in the cell, and the input gate controls the flow of new values into the cells. We used the Adam optimization function and the sigmoid activation function in the suggested model to categorize news.

TESTING THE CLASSIFICATION RESULTS

Several methods are used to test and validate trained models. For the purpose of classifying the news that is fed into the system, the outcomes are forecast and the model's accuracy is computed.

VII.IMPLEMENTATION

TECHNOLOGIES USED

Python is utilized for the im plementation of two models since, in comparison to other computer languages, it is better suited to handle Machine Learning approaches. Since the LSTM model requires TensorFlow, which in turn requires a GPU to run, Google Collab is utilized to execute and run the models. The LSTM model is executed using the GPU runtype, while the SVM model is executed using the conventional runtype. Numpy, Pandas, Seaborn, Ski-kit Learn, and Matplotlib are just a few of the libraries used for data partitioning, training, and visualization.

SVM IMPLEMENTATION

1. Import all libraries
2. Open data files by giving the file path
3. Specify data as real or fake using flag value
4. Concatenate data frames
5. Display Data
6. Perform Vectorisation and tokenisation
7. Split data
8. Train SVM algorithm
9. Calculate accuracy of the algorithm against the dataset
10. Perform Confusion matrix method
11. Calculate Precision, Recall. F1 Score, Support Values
12. Give input and find whether it is fake or real
13. End Process

LSTM IMPLEMENTATION

1. Import all libraries
2. Open data files by giving the file path
3. Specify data as real or fake using flag value
4. Concatenate data frames

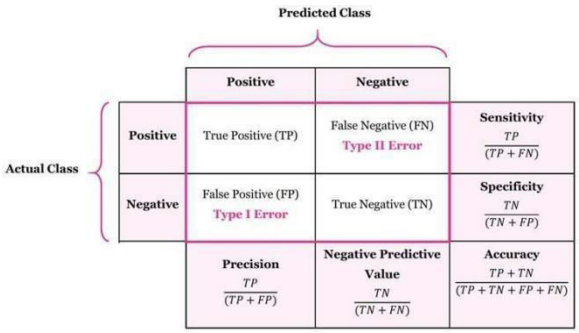
ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 08, 2022

5. Display Data
6. Perform Vectorisation and tokenisation
7. Split data
8. Train LSTM algorithm
9. Calculate accuracy of the algorithm against the dataset
10. Perform Confusion matrix method
11. Calculate Precision, Recall. F1 Score, Support Values
12. Give input and find whether it is fake or real
13. End Process

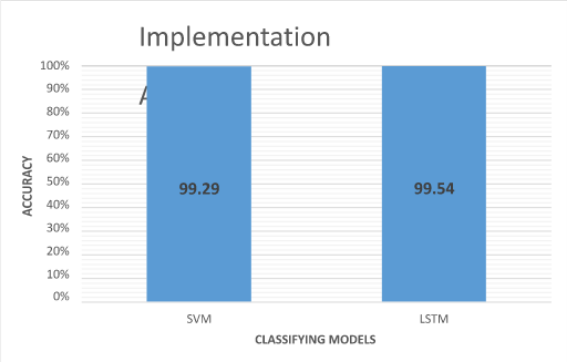
VIII. RESULT ANALYSIS

CONFUSION MATRIX

When describing the performance of a classification model (also known as a "classifier") on a set of test data for which the true values are known, a confusion matrix is a table that is frequently utilized.



PERFORMANCE ANALYSIS



RESULTS

Table 8.1: Confusion Matrix for SVM Classifiers

Test Cases	Actual Class	Predicted Class	
		Positive	Negative
Train = 80 Test = 20	Positive	4640	42
	Negative	21	4277
Train = 70 Test = 30	Positive	6998	55
	Negative	28	6389
Train = 60 Test = 40	Positive	9365	73
	Negative	45	8477
Train = 50 Test = 50	Positive	11678	101
	Negative	65	10605

Table 8.2: Confusion Matrix for LSTM Classifier

Test Cases	Actual Class	Predicted Class	
		Positive	Negative
Train = 80 Test = 20	Positive	5912	37
	Negative	51	5225
Train = 70 Test = 30	Positive	5918	31
	Negative	23	5253
Train = 60 Test = 40	Positive	5915	34
	Negative	33	5243
Train = 50 Test = 50	Positive	5919	30
	Negative	26	5250

Table 8.3:Accuracy & Classification for SVM

Test Cases	SVM Model				
Train = 80 Test= 20	**Accuracy :** 99.2984409799556				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	0.99	4682
	Real	0.99	1.00	0.99	4298
	Avg	0.99	0.99	0.99	8980
Train = 70 Test= 30	**Accuracy :** 99.38381588715664				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	0.99	7053
	Real	0.99	1.00	0.99	6417
	Avg	0.99	0.99	0.99	13470
Train = 60 Test= 40	**Accuracy :** 99.34298440979956				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	0.99	9438
	Real	0.99	0.99	0.99	8522
	Avg	0.99	0.99	0.99	17960
Train = 50 Test= 50	**Accuracy :** 99.26054612677625				
		Precision	Recall	f1-Score	support
	Fake	0.99	0.99	0.99	11779
	Real	0.99	0.99	0.99	10670
	Avg	0.99	0.99	0.99	22449

Table 8.3 shows the Long Short-Term Memory (LSTM) model under different train-test splits. The metrics used to evaluate the model include **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **Support**

Figure 8.1: SVM model performance metrics

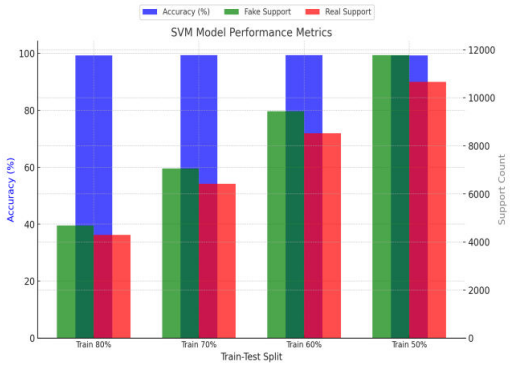


Figure 1 shows the SVM model's performance metrics. It includes:

- **Accuracy**: Shown as blue bars.
- **Fake Support**: Shown as green bars.
- **Real Support**: Shown as red bars.

The x-axis represents the different train-test splits, while the y-axes show the accuracy percentage and the support counts.

Figure 8.2: LSTM model performance metrics

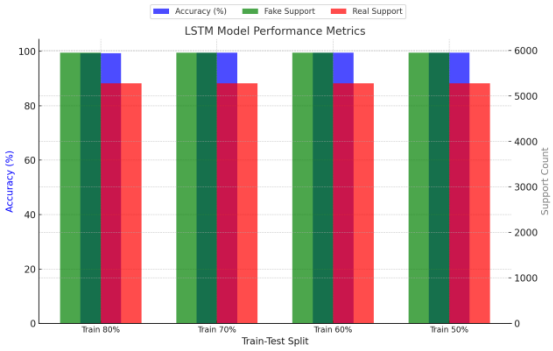


Figure 2 shows The LSTM model's performance metrics, which includes:

- **Accuracy**: Shown as blue bars.
- **Fake Support**: Shown as green bars.
- **Real Support**: Shown as red bars.

The x-axis represents different train-test splits, while the y-axes represent the accuracy percentage and support counts. Let me know if you need further customizations or comparisons with other models

Table 8.4: Accuracy & Classification for LSTM

Test Cases	LSTM Model				
Train = 80 Test= 20	**Accuracy :** 99.3103563474387				
		Precision	Recall	f1-Score	support
	Fake	0.99	0.99	0.99	5949
	Real	0.99	0.99	0.99	5276
	Avg	0.99	0.99	0.99	11225
Train = 70 Test= 30	**Accuracy :** 99.51893095768374				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	1.00	5949
	Real	0.99	1.00	0.99	5276
	Avg	1.00	1.00	1.00	11225
Train = 60 Test = 40	**Accuracy :** 99.50311804008909				
		Precision	Recall	f1-Score	support
	Fake	0.99	0.99	0.99	5949
	Real	0.99	0.99	0.99	5276
	Avg	0.99	0.99	0.99	11225
Train = 50 Test = 50	**Accuracy :** 99.5011135857461				
		Precision	Recall	f1-Score	support
	Fake	1.00	0.99	1.00	5949
	Real	0.99	1.00	0.99	5276
	Avg	1.00	1.00	1.00	11225

Table 8.4 shows the SVM model under different train-test splits. The metrics used to evaluate the model include **Accuracy**, **Precision**, **Recall**, **F1-Score**, and **Support**

IX. CONCLUSION and FUTURE SCOPE

Our suggested method uses the SVM approach (ML) and the LSTM approach (DL) to classify the supplied article as legitimate or fake news based on the information provided in that article. This makes it easier for consumers to determine if the news they get online and on social media platforms is accurate or deceptive.

The SVM classifier has an accuracy of 99.29%, whereas the LSTM classifier has an accuracy of 99.54%. While both algorithms provide decent accuracy, when it comes to identifying news articles LSTM performs better than SVM. After examining different scenarios, it's clear that the LSTM model consistently achieved better results than the SVM model in terms of accuracy. The LSTM model's classification results showed more precise values than those of the SVM model. By employing an LSTM model rather than an SVM model, the articles were more correctly identified as either fake news or genuine news. Thus, the LSTM model excels over the SVM model in identifying fake news. In the future, fake news detection can be enhanced by developing models that can identify fake news using audio, video, and image data. To create effective models, the accuracy of various fake news detection models utilizing different algorithms can be compared.

REFERENCES

[1] Sobhani, Svetlana Kiritchenko, Parinaz, and Saif Mohammad. (2016). "Detecting stance in tweets and analyzing its interaction with sentiment." The Fifth Joint Conference on Computational and Lexical Semantics: Proceedings.

[2] Parinaz and Sobhani (2017) conducted a Ph.D. dissertation titled "Stance Detection and Analysis in social media." Ottawa University.

[3] Richard Davis and Chris Proctor.ET.AL. "Fake News, Real Consequences: Recruiting Neural Networks for the Fight Against Fake News."

[4] Ayman Aljabouh, Alim Al Ahmed, Myung Suh Choi, and Praveen Kumar Donepudi. February 8, 2021. "KNN model-based machine learning classifiers for the detection of fake news," available at arXiv:2102.04458[cs.CY].

[5] In June of 2017, K. Shu, A. Sliva, S., Wang, J., Tang & H. Liu. ACM New York's newsletter, "Fake News Detection on Social Media: A Data Mining Perspective," 10.1145/3137597.3137600 can be accessed here.

[6] Chen, Y., Conroy, N.J., Rubin, and V.L. June of 2016. "Truth or false information?" "Detecting possibly misleading news using satirical cues: proceedings of the second workshop on computational approaches to deception detection." as part of NAACL-HLT.

[7] Steinberger, Krejzl, & (2016). a challenge of stance detection on tweets that was "focused on using domain knowledge related features." In Task 6 of SemEval.

[8] Aroofa Maryam, Zameer Ahmed, Nooruz-Zuha, Aroofa Nada, and Bariya Firdous Khan. (May 5, 2019). The article "Fake news detection using logistic regression" was published in the 06 issue of the International Research Journal of Engineering and Technology (IRJET).

ISSN PRINT 2319 1775 Online 2320 7876
 UGC CARE Listed (Group -I) Journal Volume 11, Iss 08, 2022

[9] Aylien Brambrick (April 20, 2019). https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html "Support Vector Machines: A Simple Explanation."

[10] In the October 1997 issue of the Institute of Bioinformatics Publications Newsletter, Hochreiter and Jrgen discussed "Long short-term member.