Analyzing Numerical Patterns in Twitter Data: Unveiling Fake and Bot Accounts During Telangana State Elections

Amrit Kumar Singha and
Arnov Paul
Department of ECE
SRM University, Amaravati
Andhra Pradesh, India
amrit_manash@srmap.edu.in
arnov_paul@srmap.edu.in

Sasank Sonti and
Karthik Guntur
Department of ECE
SRM University, Amaravati
Andhra Pradesh, India
sasank_sonti@srmap.edu.in
karthik_guntur@srmap.edu.in

Mondikathi Chiranjeevi and
Sateeshkrishna Dhuli
Department of ECE
SRM University, Amaravati
Andhra Pradesh, India
chiranjeevi_m@srmap.edu.in
sateeshkrishna.d@srmap.edu.in

***Abstract*—Online Social Networks (OSNs) have become ubiquitous platforms for the dissemination of diverse content, encompassing text, images, and videos. However, the proliferation of fake accounts poses a formidable challenge to the integrity of current OSN systems. Exploiting these fraudulent profiles, malicious actors distribute misleading information, ranging from deceptive surveys to fabricated reports of election rigging and false narratives about the government. Our proposal involves utilizing the latest developments in deep learning, namely in computer vision to address the widespread problem of phony accounts. namely, we suggest implementing an Artificial Neural Network (ANN) and certain Machine Learning methods. Through a series of experiments, our findings reveal a promising outcome, demonstrating superior accuracy and minimal loss when compared to prevalent learning algorithms in the realm of fake account classification. This research contributes to the ongoing discourse on enhancing the robustness of OSN systems against the propagation of misleading information through fake accounts.**

***Index Terms*—Fake Profiles, Social media, Twitter, Machine Learning, Numeric data Preprocessing, Feature extraction, Model evaluations, State Election, Neural Networks, Multi-Layer perception.**

I. Introduction

The widespread adoption of advanced technologies and the ubiquitous availability of Internet access have fuelled the exponential growth of online social networking (OSN) websites, captivating a vast global audience. Among the most well-known OSN platforms are Facebook (FB), Twitter, Instagram, YouTube, and LinkedIn. These platforms have an astounding number of registered users from all over the world [1]. These platforms provide a cost-free and easily accessible means for individuals to communicate with their connections about updates, activities, and interests that are either personal or professional. They make it easier for linked users to easily communicate messages, images, videos, online diaries (blogs), and other types of content [2]. In today's interconnected world, social media has emerged as an indispensable tool for communication, business promotion, and marketing [1].

Fundamentally, OSN websites are interactive computer-mediated technologies that facilitate the development, sharing, and administration of knowledge, concepts, career goals, and various kinds of expression via online communities and networks [3]. Another popular OSN platform, Twitter, boasts 372.9 million monthly active users. YouTube and Instagram exhibit similar levels of user engagement. A significant advantage of OSN websites is their ability to bridge the gap between individuals of all ages and backgrounds, enabling them to stay connected with old and new friends. These platforms also facilitate the formation of new connections with like-minded individuals across the globe [5]. These capabilities, coupled with the myriad ways to communicate with friends, make OSN websites highly appealing to users. For many, their social lives, and even their practical lives, have become inextricably intertwined with these online communities[6]. Despite their immense popularity, OSN websites face a multitude of challenges, including security concerns, privacy breaches, spamming, rumour-mongering, and the proliferation of fake profiles. The lack of restrictions on profile creation on OSN websites has provided unscrupulous individuals with ample opportunities to establish fake profiles and exploit the platform for personal or organizational gains. Popular OSN websites like Twitter and FB are actively addressing these issues by proactively identifying and removing fake profiles [1][2].

The pervasiveness of unethical, untruthful, rumour-mongering, and spam messages, often disseminated through fake accounts or profiles, has become a prevalent issue on online social networking (OSN) websites [4]. While significant research efforts have been directed towards detecting spam messages, the identification of fake profiles remains a formidable challenge. Fake profiles are often created by exploiting readily available data from current profiles on the internet, such as profile names, profile photos, age, gender, and other personal information. This misuse of personal data exposes incorrect and potentially misleading information to friends and contacts interconnected through social media [2].

The consequences of such fabricated information can extend far beyond the online realm, potentially causing significant harm to individuals, businesses, and society as a whole. Fake accounts or profiles can be used to spread fake news, manipulate online ratings, and promote spam content. OSN operators are currently dedicating substantial resources to detect, verify, and eliminate fake profiles[1].

Fig.1: Statistics of news reaching through online. The fig.1 shows that 75% of people read all news online. Of these, 70% read election news online. 57% of people read news websites, and 41% read social media. This suggests that online news is the most popular way to read election news. News websites are more popular than social media for reading election news. The fig.1 also shows that there is a significant overlap between people who read all news online and people who read election news online [7]. This suggests that people who are interested in news in general are also likely to be interested in election news. From this we can conclude that elections will have a greater impact from OSN and misinformation may lead to an imbalance in democracy.

II. Dataset

The training dataset comprises 22 features, including tweet details such as text content, creation timestamp, retweet and favorite counts, user information like ID, screen name, and followers count. The testing dataset, with 22 features as well, includes language information, tweet metrics, user identifiers, and tweet-related counts. Both datasets share common features like text content and user-related information, essential for training and evaluating models. Notably, the training dataset consists of 250 samples, while the testing dataset comprises of 2049 samples.

Fig. 1. Distribution Of Data

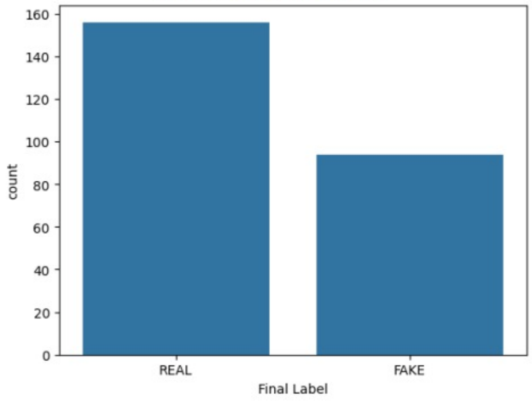

TABLE I
Features of the datasets

training data	testing data
Unique ID	Lang
Text	Edit_history_tweets_ids
created_at	Public_metrics
Retweet_count	Created_at
Favorite_count	Author_id
Source	Id
length	Possibly_sensitive
User_id	Text
User_screen_name	Retweet_count
User_name	Reply_count
User_created_at	Like_count
User_description	Quote_count
User_followers_count	Bookmark_count
User_friends_count	Impression_count
User_location	Profile_image_url
User_statuses_count	Username
User_verified	Description
User_url	Id.1
Statuses/follorers_count	Name
Friends/followers_count	Verified
User_has_url?	Verified_type
Final Label	User has profile pic

III. Methodology

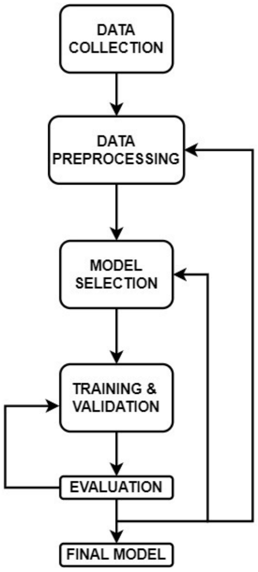

Fig. 2. Flowchart

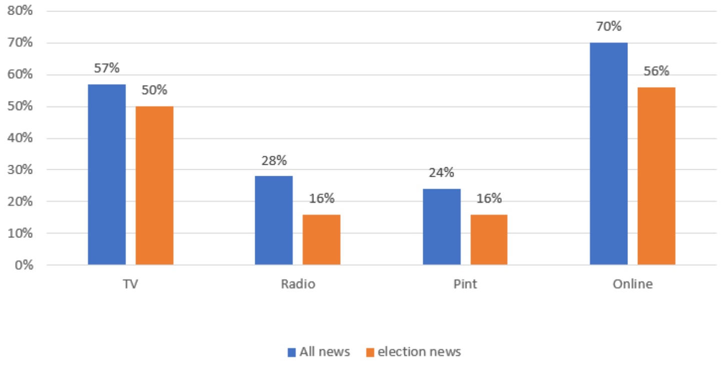

Fig. 1. Statistics of news reaching through online

To proceed further in finding the originality of an OSN account, what we propose is to perform numerical analysis on the data set by taking the features such as Likes Count, Retweet Count, User Verified and Profile Pic of the data set into account as shown in Fig. II.

A. *Learning Models*

*1) **Logistic regression**:* Logistic regression is a fundamental statistical technique widely used in various fields for both prediction and classification tasks. It excels in modelling the probability of an event occurring, particularly when dependent on one or more independent variables. Unlike linear regression, which predicts continuous values, logistic regression outputs estimates between 0 and 1, representing the probability of belonging to a specific category [8].

Mathematical Framework:

Logistic regression relies on the sigmoid function (also known as the logistic function), which maps any real number to a value between 0 and 1. This function forms the basis of the model, relating the independent variables (X) to the probability of the event occurring (P):

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n)}} \tag{1}$$

Here, the β coefficients represent the model parameters, quantifying the influence of each independent variable (X) on the probability of the event (y = 1). The intercept β_0 accounts for the baseline probability even when all independent variables are zero [9].

*2) **AdaBoost**:* Adaptive Boosting (AdaBoost) is a prominent ensemble learning algorithm in machine learning that excels in classification tasks. It achieves high accuracy by sequentially combining multiple ”weak learners” into a powerful ”strong learner.” This feature makes it well-suited for scenarios where individual base learners are insufficiently discriminative, but their collective knowledge can lead to a robust model [10].

Mathematical Framework:

AdaBoost iteratively builds an ensemble of weak learners, with each iteration focusing on examples misclassified by previous learners. Formally, let:

- D_t be the training data distribution at iteration t.
- w_ti be the weight assigned to instance i at iteration t (initially uniform).
- h_t be the weak learner chosen at iteration t.
- α_t be the boosting coefficient for h_t [11].

*3) **Voting Classifier**:* Voting classifiers, sometimes referred to as ensemble classifiers, are potent machine learning algorithms that enhance overall classification resilience and accuracy by combining the predictions of several base learners. When compared to single-learner models, they perform better because they maximize each learner’s strengths while reducing their limitations [12].

A **Theoretical Structure:**

Voting classifiers work by combining the predictions of several base learners, which are usually selected from different kinds of models. After doing its own independent analysis of the data, each base learner generates a prediction for the target class. These separate forecasts are combined using a voting system to generate the final projection, such as

- Majority Voting: The most frequently predicted class among all base learners becomes the final prediction [13].
- Weighted Voting: Assigns weights to each base learner based on their individual performance or confidence levels, with the weighted sum of predictions determining the final class [14].
- Mean: This is applicable to continuous target variables, where the final prediction is obtained by averaging the individual predictions made by base learners [23].

*4) **Artificial Neural Network (ANN)**:* An ANN is a type of computational model that draws inspiration from the architecture and mechanisms of the human brain. It consists of interconnected layers of artificial neurons, resembling the network of biological neurons in the brain. Each neuron receives inputs from other neurons, processes them using an activation function, and transmits an output signal to other neurons. This interconnected web of neurons allows ANNs to learn and adapt to complex patterns in data, similar to how the human brain learns through experience [16], [17].

The weights of the connections between neurons are changed by ANNs in order to learn. Backpropagation is the technique of adjusting the weights in a network by comparing the output of the network to the desired output and propagating the error back through the network. Common learning algorithms include:

- **Gradient descent:** This algorithm iteratively updates the weights in the direction that minimizes the error between the network's output and the desired output [18].
- **Stochastic gradient descent:** A more efficient version of gradient descent that updates weights based on a subset of the training data instead of the entire dataset [19].
- **Adam:** A variant of gradient descent that adapts the learning rate for each parameter, leading to faster convergence and improved performance [20].

5) Random Forest Classifier (RF): An ensemble of decision trees, each modeling the target variable independently using a collection of characteristics and split points selected at random, is called a regression forest (RF). Every tree expands throughout training without being pruned, enhancing the amount of information gained at every node. By combining the predictions of each individual tree, the RF's final prediction is produced. This is usually done by averaging the regression predictions or using majority vote for categorization. When compared to single decision trees, the model's overall resilience and accuracy are increased by this ”wisdom of the crowds” notion [25].

Random forests consist of two key elements:

- **Decision Trees:** These are tree-like structures where each node represents a decision on a feature, leading to branches based on the decision outcome. Leaves at the end of the tree represent predictions for the target variable [22].
- **Ensemble:** Many such decision trees are combined in the forest, with each tree voting on the final prediction. The majority vote determines the final prediction of the random forest [15].

6) Gradient Boosting Classifier (GB): Using an ensemble of weak prediction models—usually decision trees—gradient boosting is a potent machine learning technique that iteratively constructs a stronger and more accurate model. This method's adaption for classification applications is called the gradient boosting classifier [29].

7) Multi Layer Perception (MLP): An artificial neural network consisting of interconnected layers of artificial neurons is called an MLP. Signals are propagated from the input layer via hidden layers and ultimately to the output layer by the network as it processes information. Every neuron calculates its output by applying an activation function to the weighted sum of its inputs [21].

Learning Algorithm:

MLPs make use of backpropagation, a training process that modifies link weights iteratively in response to variations between the expected and predicted outcomes of the network. By optimizing the weights, the error is reduced and the accuracy of the network is enhanced [26].

8) XG Boosting: Gradient boosting is a technique that creates an ensemble of weak learners (decision trees) iteratively to enhance prediction accuracy. XGBoost is an optimized version of this technique. Regularization and loss function minimization are used to keep the model from overfitting and improve its performance [27].

XGBoost builds an ensemble of decision trees where each tree:

- Learns from weighted residuals of previous trees to focus on areas with higher errors.
- Splits features based on information gain, maximizing predictive power with each step.
- Has its contribution to the final prediction controlled by its performance and complexity [27].

9) Hyperparameter Tuning Using Grid Search: A methodical technique for fine-tuning hyperparameters in machine learning models is grid search. It entails creating a grid of potential hyperparameter values, thoroughly testing the model with every possible combination, and determining which combination of parameters produces the best results [28].

Key Features:

- **Exhaustive search:** Grid search systematically explores all possible combinations of hyperparameter values within the predefined grid, increasing the likelihood of finding the optimal configuration.
- **Simplicity:** Researchers of all skill levels can use the notion because it is simple to understand and apply.
- **Effectiveness:** Grid search often leads to significant improvements in model performance, especially for models sensitive to hyperparameter settings [24].

IV. Results

After testing the mentioned models, we got the results as shown in the Table.2.

TABLE II
Results of the models

Models	Precision	Recall	F1-score	Accuracy
Logistic Regression	0.62	0.94	0.75	0.60
AdaBoost	0.66	0.97	0.78	0.66
Voting Classifier	0.63	0.97	0.77	0.62
ANN	0.65	0.69	0.67	0.56
RF Classifier	0.66	0.97	0.78	0.66
GB Classifier	0.68	0.66	0.67	0.58
MLP	0.64	1.00	0.78	0.64
XG Boosting	0.65	1.00	0.79	0.66
Grid Search	0.66	0.97	0.78	0.66

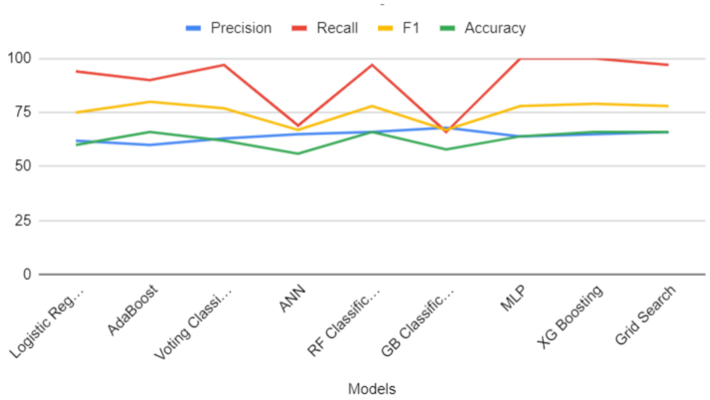

Fig. 3. Comparison graph of the models

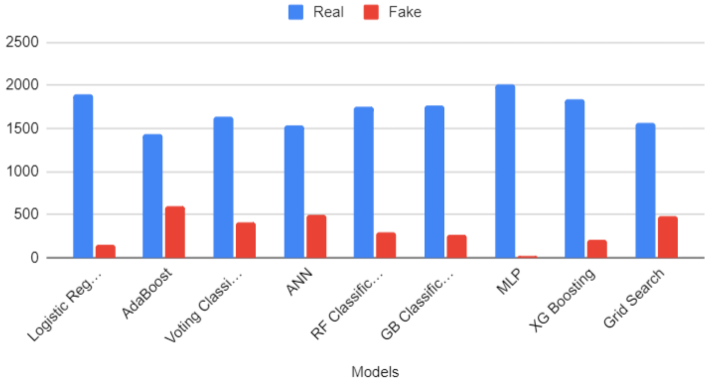

Fig. 4. Graph of real and fake accounts detected

TABLE III
REAL AND FAKE ACCOUNTS DETECTED BY EACH MODEL

Models	Real	Fake
Logistic Regression	1894	154
AdaBoost	1444	604
Voting Classifier	1633	415
ANN	1543	505
RF Classifier	1752	296
GB Classifier	1773	275
MLP	2020	28
XG Boosting	1837	211
Grid Search	1567	481

We can do the cross-validation of the results by calculating precision, recall, f1-score and accuracy by the following formulas [30]:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

V. CONCLUSIONS

In conclusion, our study has yielded valuable insights into the multifaceted realm of Twitter discourse during the Telangana State Elections. Despite leveraging a diverse range of machine learning models, the quest for optimal accuracy proved to be a nuanced endeavor.

Through meticulous analysis, underpinned by metrics such as likes and retweets, we discerned patterns suggestive of the elusive nature of digital misrepresentation. The interplay between precision and recall provided illuminating perspectives on the challenges inherent in preserving authentic dialogue amidst the identification of deceptive entities.

It is imperative to acknowledge the dynamic nature of counterfeit accounts and the inherent limitations in formulating universally applicable solutions. The pursuit of accuracy remains an ongoing imperative, warranting continued exploration and refinement.

In sum, our research contributes to a nuanced comprehension of Twitter's electoral discourse within the context of Telangana. As the reverberations of political discourse persist, we extend an invitation to future scholars to join us in navigating the ever-evolving landscape of social media authenticity.

REFERENCES

[1] Wanda, Putra, and Huang Jin Jie. ”DeepProfile: Finding fake profile in online social network using dynamic CNN.” Journal of Information Security and Applications 52 (2020): 102465.

[2] Roy, Pradeep Kumar, and Shivam Chahar. ”Fake profile detection on social networking websites: a comprehensive review.” IEEE Transactions on Artificial Intelligence 1.3 (2020): 271-285.

[3] Shahane, P. R. I. Y. A. N. K. A., and D. E. I. P. A. L. I. Gore. ”Detection of fake profiles on Twitter using random forest & deep convolutional neural network.” Int. J. Manag. Technol. Eng 9 (2019): 3663-3667.

[4] Homsi, Ahmad, et al. ”Detecting Twitter Fake Accounts using Machine Learning and Data Reduction Techniques.” DATA. 2021.

[5] Rahman, M. D., et al. Detection of fake identities on Twitter using supervised machine learning. Diss. Brac University, 2019.

[6] Ramalingaiah, A., S. Hussaini, and S. Chaudhari. ”Twitter bot detection using supervised machine learning.” Journal of Physics: Conference Series. Vol. 1950. No. 1. IOP Publishing, 2021.

[7] Digital News Report: A Mile Wide, an Inch Deep: Online News and Media Use in the 2019 UK General Election at https://www.digitalnewsreport.org/publications/2020/mile-wide-inch-deep-online-news-media-use-2019-uk-general-election/attachment/slide2_opt/

[8] Landwehr, Niels, Mark Hall, and Eibe Frank. ”Logistic model trees.” Machine learning 59 (2005): 161-205.

[9] Jordan, Michael I. ”Why the logistic function? A tutorial discussion on probabilities and neural networks.” (1995).

[10] Rojas, Raúl. ”AdaBoost and the super bowl of classifiers a tutorial introduction to adaptive boosting.” Freie University, Berlin, Tech. Rep 1.1 (2009): 1-6.

[11] Ferreira, Artur J., and Mário AT Figueiredo. ”Boosting algorithms: A review of methods, theory, and applications.” Ensemble machine learning: Methods and applications (2012): 35-85.

[12] Mahabub, Atik. ”A robust technique of fake news detection using Ensemble Voting Classifier and comparison with other classifiers.” SN Applied Sciences 2.4 (2020): 525.

[13] Gandhi, Isha, and Mrinal Pandey. ”Hybrid ensemble of classifiers using voting.” 2015 international conference on green computing and Internet of Things (ICGCIoT). IEEE, 2015.

[14] Kuncheva, Ludmila I., and Juan J. Rodríguez. ”A weighted voting framework for classifiers ensembles.” Knowledge and information systems 38 (2014): 259-275.

[15] Dietterich, Thomas G. ”Ensemble methods in machine learning.” International workshop on multiple classifier systems. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000.

[16] Agatonovic-Kustrin, S., and Rosemary Beresford. ”Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research.” Journal of pharmaceutical and biomedical analysis 22.5 (2000): 717-727.

[17] Chen, Wun-Hwa, Sheng-Hsun Hsu, and Hwang-Pin Shen. ”Application of SVM and ANN for intrusion detection.” Computers & Operations Research 32.10 (2005): 2617-2634.

[18] Baldi, Pierre. ”Gradient descent learning algorithm overview: A general dynamical systems perspective.” IEEE Transactions on neural networks 6.1 (1995): 182-195.

[19] Yacouby, Reda, and Dustin Axman. ”Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models.” Proceedings of the first workshop on evaluation and comparison of NLP systems. 2020.

[20] Tong, Qianqian, Guannan Liang, and Jinbo Bi. ”Calibrating the adaptive learning rate to improve convergence of ADAM.” Neurocomputing 481 (2022): 333-356.

[21] Parmar, Aakash, Rakesh Katariya, and Vatsal Patel. ”A review on random forest: An ensemble classifier.” International conference on intelligent data communication technologies and internet of things (ICICI) 2018. Springer International Publishing, 2019.Popescu, Marius-Constantin, et al. ”Multilayer perceptron and neural networks.” WSEAS Transactions on Circuits and Systems 8.7 (2009): 579-588.

[22] Charbuty, Bahzad, and Adnan Abdulazeez. ”Classification based on decision tree algorithm for machine learning.” Journal of Applied Science and Technology Trends 2.01 (2021): 20-28.

[23] Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. ”Effective voting of heterogeneous classifiers.” European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[24] Alibrahim, Hussain, and Simone A. Ludwig. ”Hyperparameter optimization: Comparing genetic algorithm against grid search and bayesian optimization.” 2021 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2021.

[25] Parmar, Aakash, Rakesh Katariya, and Vatsal Patel. ”A review on random forest: An ensemble classifier.” International conference on intelligent data communication technologies and internet of things (ICICI) 2018. Springer International Publishing, 2019.

[26] Almeida, Luis B. ”Multilayer perceptrons.” Handbook of Neural Computation. CRC Press, 2020. C1-2.

[27] Nalluri, Mounika, Mounika Pentela, and Nageswara Rao Eluri. ”A Scalable Tree Boosting System: XG Boost.” Int. J. Res. Stud. Sci. Eng. Technol 7 (2020): 36-51.

[28] Shekar, B. H., and Guesh Dagnew. ”Grid search-based hyperparameter tuning and classification of microarray cancer data.” 2019 second international conference on advanced computational and communication paradigms (ICACCP). IEEE, 2019.

[29] Ali, Zeravan Arif, et al. ”Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review.” Academic Journal of Nawroz University 12.2 (2023): 320-334.

[30] Ketkar, Nikhil, and Nikhil Ketkar. ”Stochastic gradient descent.” Deep learning with Python: A hands-on introduction (2017): 113-132.