Fake Job Post Detection Using Machine Learning

1st Pradeep Kumar B P
Dept. of Computer Science and Engineering
Atria Institute of Technology
Bangalore, Karnataka, Indiapradi14cta@gmail.com

2nd Bhagya D J
Dept. of Computer Science and Design
Atria Institute of Technology
Bangalore, Karnataka, India
bhagyadjbhagyadj@gmail.com

3rd Gagana G
Dept. of Computer Science and Design
Atria Institute of Technology
Bangalore, Karnataka, India
gagana503675@gmail.com

4th Mani N
Dept. of Computer Science and Design
Atria Institute of Technology
Bangalore, Karnataka, India
manin982001@gmail.com.

2025 International Conference on Computing for Sustainability and Intelligent Future (COMP-SIF) | 979-8-3315-3853-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/COMP-SIF65618.2025.10969867

***Abstract*-By examining patterns in the text data and metadata linked to job adverts, The use of learning algorithms has grown in popularity in the detection of fraudulent job postings. For this, Research Machine emphasises the application of a number of algorithms, such as Naive Bayes and Random Forest , Together with models for deep learning, like CNN and Bidirectional LSTM, and AdaBoost. To extract useful features These techniques depend on text preparation techniques such as lemmatisation, tokenisation, and TF-IDF vectorisation. To differentiate between authentic and fraudulent posts, models are instructed in datasets and evaluated using measures such as accuracy, accuracy, memory, and F1-score. Advanced classifiers and group techniques have proven to be quite successful in enhancing detection performance while lowering false positives and negatives.**

***KeyWords:* Fake Job Ads Detection, Models for Deep Learning, Naive Bayes, TF-IDF Vectorisation, Ensemble Techniques**

I.INTRODUCTION:

Because online recruitment platforms are so popular in the digital age, fake job advertisements have grown to be a serious problem. These scams frequently aim to deceive job seekers into disclosing personal information[21]. or paying money, have detrimental effects that include financial losses, psychological distress, and harm to the standing of reputable companies and employment boards. The intricacy and scale of these fraudulent operations have caused a boom in the study of automated detection systems based on artificial intelligence and machine learning [22].In order to differentiate between genuine and fraudulent job advertisements, these algorithms examine the intrinsic patterns and characteristics of these ads [23].

Using both conventional and cutting-edge techniques, machine learning provides an efficient framework for addressing this issue [24]. Techniques like logistic regression, Naive Bayes, and Support Vector Machines (SVM) have been studied for classification problems because to their effectiveness and user-friendliness [25].By pooling the predictions of several models, ensemble techniques such as Random Forest, Gradient Boosting, and AdaBoost have shown greater performance, increasing accuracy and resilience.Two recent advancements in deep learning architectures that excel at processing and understanding the textual data found in job descriptions are Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks.

The success of these models primarily depends on data pretreatment and feature engineering. Techniques like tokenisation, lemmatisation, and Inverse Document Frequency-Term Frequency (TF-IDF) vectorisation are commonly used to transform unprocessed textual input into a machine learning-ready format. Feature selection procedures further enhance model performance by cutting down on noise and computing cost by removing the dataset's most pertinent features. Numerous performance metrics are employed by researchers, including F1-score, recall, accuracy, and precision. and Cohen-Kappa scores, to assess how well these detection algorithms work.

II. LITERATURE REVIEW:

The 20 papers that are linked examine a variety of machine learning, deep learning, and data preparation modalities, benefits, and applications. Bidirectional LSTM, CNNs, U-Net architectures, Random Forest and other ensemble classifiers and AdaBoost, and text processing techniques like TF-IDF and tokenisation are among the modalities used in these works. By comprehending the context of text data, bidirectional LSTM, for example, is well-suited to identifying fraudulent job posts because of its exceptional efficacy in processing sequential data. Conversely, ensemble classifiers use the combined power of several models to increase robustness and accuracy, particularly in noisy datasets. Similar to this, CNNs and U-Net architectures perform exceptionally well in image-based assignments because of their capacity to capture spatial and hierarchical features, which makes them very helpful for the segmentation and improvement of medical images.These modalities' efficacy and applicability clearly demonstrate their benefits. For sequential text data, bidirectional LSTM provides high detection accuracy by capturing both past and future context. When it comes to

to feature selection, group methods such as Forest at Random and Gradient Boosting are reliable and efficient, improving classification accuracy. Machine learning classifier performance is greatly increased by text processing methods like TF-IDF and lemmatisation, which enhance data representation and lower noise. In contrast, CNN and U-Net designs offer excellent accuracy in image segmentation tasks, especially for medical applications like histopathological imaging and cancer detection. Depending on the objective, these approaches show a balance between resilience, adaptability, and efficiency.

These methods have a wide range of useful applications. Techniques like ensemble classifiers and bidirectional LSTM have been used in fraudulent job identification to reliably identify phoney job listings. These classifiers are enhanced by text pretreatment techniques, which guarantee that the data is clear and accurately represented. U-Net and Wavelet Transform for Dual Tree Complex have been used in the healthcare industry for tasks like cancer diagnosis, medical picture segmentation, and imaging quality improvement. Another important application field is gesture recognition, where real-time sign language identification and gesture detection are made possible by techniques like CNNs, Bag of Features, and Kinect-based recognition. Together, these research demonstrate the adaptability and potency of deep learning and machine learning approaches across fields, providing answers to urgent problems in fraud detection, healthcare, and human-computer interaction

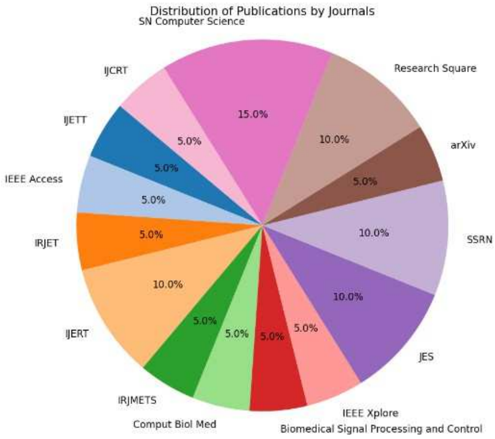


Figure 1.Publication Distribution by Journals

The First Figure, A pie chart displays the distribution of publications among several periodicals. Each slice of the pie represents a journal, and the size of the slice indicates the proportion of all publications that can be attributed to that journal. On the chart, each notebook is represented by a different color, which makes it easy to visually differentiate between themThe precise contribution of each journal to the overall number of articles is indicated by the percentage labels inside each slice. With larger slices denoting journals with more publications, this visualisation offers insights into which magazines are most commonly cited in the dataset.

*The Second Figure***,** shows the distribution of publications throughout several years in a bar chart. The y-axis shows the number of publications in each year, while the x-axis shows the years, such as 2021, 2022, 2023, and so forth. Each bar's height represents the total number of publications released during that specific year.

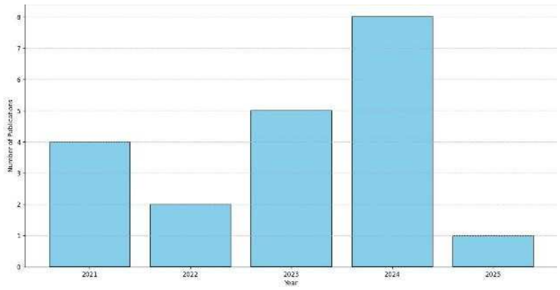


Figure 2. Distribution of Publications by Year

For ease of comparing publication counts over years, the chart features horizontal gridlines and a light blue colour scheme with black borders surrounding each bar.

Trade off Parameter: The Trade off parameters in the models covered in the table show how accuracy, interpretability, computing efficiency, and data requirements are all balanced. High precision models, including CNNs, Random Forest, and Bidirectional LSTM, are less suitable for real-time applications unless there is a large amount of hardware available because they often require a lot of processing power and training time. However, while simpler models like Naive Bayes or Logistic Regression are faster to develop and require less computing power, they might not be capable of recognise complex patterns in the data, applications like fraud detection where it's vital to understand how the model makes decisions. Simpler models, on the other hand, are easier to understand but could have trouble with more difficult jobs. Furthermore, speedier models may compromise robustness in noisy situations, while ensemble methods are computationally costly and might take longer to train, despite their high accuracy and resilience to noisy data.

Research Gap Identification: Regarding research gaps, a number of topics should be investigated to improve these models' efficacy. Improving the models' performance on little datasets is a major gap It would be beneficial to create strategies or models that work well. with little information, as many sophisticated models, such as CNN or LSTM, require big datasets for optimal performance. The interpretability of deep learning models is another weakness; Even if these models are highly precise, their opaqueness is a serious problem, particularly in applications like fraud detection,which restricts their implementation in real-time scenarios. This difficulty might be solved by investigating ways to optimise existing models for quicker inference times or creating lightweight models that maintain accuracy without using a significant amount of computing power. Furthermore, the majority of current models are quite specialised for particular purposes like fraud detection or phoney job postings. It is possible to develop general-purpose models that require little fine-tuning and may be readily applied to a range of domains. Additionally, hybrid models that integrate the advantages of several methods might provide more trustworthy outcomes. Finally, but just as importantly, even though a lot of models depend on manual feature engineering and preprocessing There is an opportunity to investigate automated feature extraction techniques

Citation No	Modality	Advantages	Applications
[1]	Bidirectional LSTM	Captures sequential data effectively, high detection accuracy for fake job postings	Fake job posting detection
[2]	Ensemble Classifiers (Random Forest, AdaBoost, Naive Bayes)	High accuracy with ensemble methods, robust to noisy data	Job recruitment fraud detection
[3]	Random Forest	Simple to implement, performs well on structured data	Fraudulent job detection
[4]	Multi-Layer Perceptron (MLP)	Flexible architecture, good for complex data	Fake job post classification
[5]	Logistic Regression, Naive Bayes	Easy to implement, interpretable	General-purpose fake job detection
[6]	Gradient Boosting, AdaBoost	High accuracy, effective feature selection	Scam detection in job postings
[7]	Decision Trees, Random Forest	Easy to interpret, handles categorical data well	Scam classification
[8]	Text Processing (TF-IDF, Lemmatization)	Enhances classifier performance, reduces noise	Text-based job scam detection
[9]	LSTM, CNN	Effective for unstructured text.	Fraudulent recruitment detection
[10]	Support Vector Machines, Decision Trees	Works well on smaller datasets,	Fake job classification
[11]	U-Net, U-Net+ Architectures	High accuracy for image segmentation	Medical image segment -ation
[12]	Dual Tree Complex Wavelet Transform	Preserves details In imaging, robust to noise	Medical imaging enhance -ment
[13]	VBIR with CLAHE & GLCM	Enhances image contrast,	Histopathology and cancer detection
[14]	Logistic Regression	Simple and interpretable	Basic fraud detection
[15]	Bag of Features	Effective for image recognition tasks	Sign language recognition
[16]	Support Vector Machines	Works well on smaller datasets,	Fake job classification
[17]	KNN, SVM, ANN	Versatile and adaptable to different data types	Gesture recognition systems
[18]	CNNs	Highly effective for image data	Image-based applications
[19]	Feature Engineering (TF-IDF, Tokenization)	Improves classification accuracy	Text classification
[20]	Kinect Sensor-based Recognition	Real-time recognition, user-friendly	American Sign Language recognition

or self-adaptive models that can dynamically adapt to extract the best features for a particular dataset.

III. METHODOLOGY OVERVIEW:

3.1 Data Collection: The initial phase in the fake job post prediction system's technique is data collecting, which involves gathering a broad range of job listings from multiple internet platforms like Kaggle. Important details like job title, location, pay, employer, necessary experience, and job description are all included in this dataset.

*3.2 Hardware Requirements :*A Pentium i3 processor, 500 GB of hard drive space, 2 GB of RAM, and a 15-inch LED monitor are among the technical specifications for this machine. The Python programming language, which offers the required libraries and frameworks for creating and assessing machine learning models.

*3.3 Machine Learning Models:*The system is then trained using various machine learning models after the data has been preprocessed. K-Nearest Neighbours (KNN) and Random Forest are popular models for this use. The Random Forest method is an ensemble technique that builds many decision trees and aggregates their predictions to improve accuracy and reduce overfitting.It creates variation among the trees by employing methods such as bagging.

3.4 Deep Neural Network (DNN): The task of predicting bogus job posts can also be handled by deep using these machine learning methods with learning models. The Deep Neural Network (DNN) is a well-liked option for deep learning in text-related tasks. The DNN is made up of several layers of neurones, each of which uses the data to learn progressively more complex attributes.

3.5 Block Diagram: The workflow of a Fake Job Post Detection System is represented by the block diagram, which shows the sequential steps required to identify phoney job listings. The first step in the process is data collection, which involves gathering pertinent information about job postings from a variety of sources, including curated datasets like Kaggle or web platforms..

*3.6 Mathematical Modeling:*The following mathematical models and methods are commonly employed using machine learning to identify fraudulent job posts:

Random Forest: Random Forest is an ensemble learning technique based on decision trees. It uses the following process:

Feature Selection: Choose "k" features at random from a total of "m" features.
Node Splitting: Determine the optimal split point for every feature that has been chosen.
Recursive Tree Building: Create decision trees by dividing the nodes in a certain way over and over again until you achieve a certain number of trees (n).

3.6.1 K-Nearest Neighbors (KNN):
The similarity principle is the foundation of the KNN classification method. It finds the "k" closest posts in the feature space for every new job ad, then uses the majority vote to determine the class:
Distance Metric: The Euclidean distance is commonly used to determine how similar two examples are to one another.

$$d(x, y)=\sqrt{\sum_{i=1}^{n}}(xi - yi)^2 \qquad (1)$$

Classification: The majority class among the "k" nearest neighbours is used to decide the class of a new job posting**.**

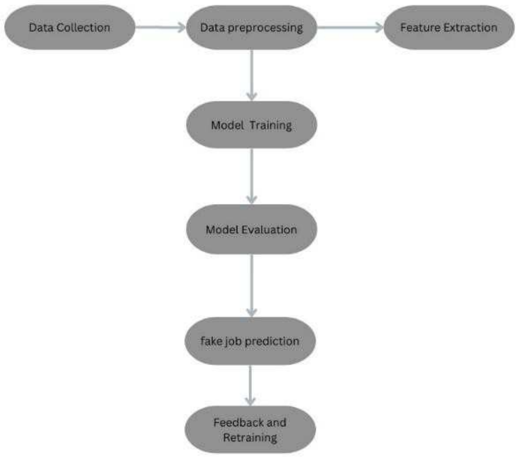


Figure3.Fake Job Post Detection System

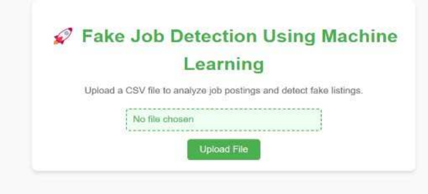
3.6.2 Algorithm:

Data Collection: Compile a labelled dataset of legitimate and fraudulent job postings.

- Preprocessing: Create an organised format out of the dataset. Standardise or normalise features. Deal with outliers and missing values.To train the model, separate the data into training and testing sets. Use the training data to train the Random Forest model.As an alternative, train the KNN model using the same data.
- Prediction: Apply the trained model (KNN or Random Forest) to new job post data to determine if the post is authentic or fraudulent.
- Model Evaluation**:** Using the testing set's Evaluate the model's performance using metrics like accuracy, precision, recall, and F1-score.
- Feedback: To help the model get better and more accurate over time, provide feedback based on its predictions.

IV. *RESULT:*

The Fig 4 ,depicts a web interface made to use machine learning to identify phoney employment advertisements. In order to detect fake job postings, the interface asks users to provide a CSV file with the job advertisements. The title, "Fake Job Detection Using Machine Learning," is followed by a succinct explanation of the tool's function. The interface features a file upload section with the wording "No file chosen," which is a placeholder, and a "Upload File" button that is clearly visible for choosing and uploading the necessary CSV file. In order to process job data and differentiate between real and fraudulent job postings, this program seems to use machine learning techniques.

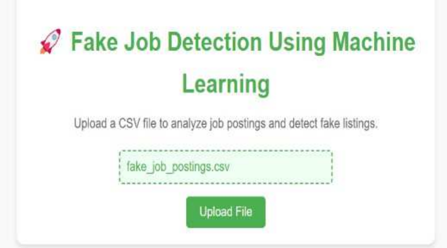


Fake Job Detection Using Machine Learning

Upload a CSV file to analyze job postings and detect fake listings.

Figure 4 Fake Job Detection Using Machine Learning Interface

The Fig 5, The user interface of a machine learning-based system intended to identify fraudulent job posts is depicted in the attached image. Users are prompted via the UI to upload a CSV file with data relevant to their jobs for analysis. The upload field displays a sample file called fake_job_postings.csv, which shows the format that the input data should take. After the file is uploaded, the system uses machine learning models to process the data and find phoney job postings.



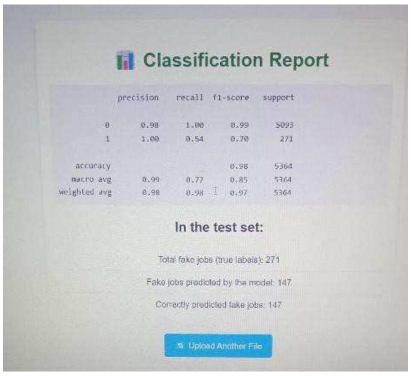
Fake Job Detection Using Machine Learning

Upload a CSV file to analyze job postings and detect fake listings.

fake_job_postings.csv

Figure 5 Fake Job Detection

The Fig 6, A classification report that summarises the effectiveness of a machine learning model for identifying fraudulent job posts is displayed in the attached image. With a 98% overall accuracy, the model performs well in detecting real jobs (precision: 98%, recall: 100%) but only moderately well in identifying phoney jobs (precision: 100%, recall: 54%). The recall for detecting phoney employment is impacted by the unequal distribution of classes, where there are much more legitimate jobs than fake ones. There is space for improvement in how the minority class is handled, as the F1-score for real jobs is 0.99 but it falls to 0.70 for phoney jobs.



Classification Report

```
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      5093
           1       1.00      0.54      0.70       271

    accuracy                           0.98      5364
   macro avg       0.99      0.77      0.85      5364
weighted avg       0.98      0.98      0.97      5364
```

In the test set:

Total fake jobs (true labels): 271

Fake jobs predicted by the model: 147

Correctly predicted fake jobs: 147

Figure 6 Classification Report

V. CONCLUSION:

In today's digital employment market, spotting phoney job postings is essential since fraudulent activities are increasingly targeting online recruitment platforms. Machine learning has become a potent instrument to deal with this problem.offering scalable and dependable techniques for identifying and removing fraudulent job advertisements. Traditional machine learning techniques like logistic regression, SVM, and Naive Bayes provide a strong basis for classification tasks, but they have trouble with complex data and unbalanced classes. Ensemble approaches like as Random Forest, AdaBoost, and Gradient Boosting significantly improve accuracy and recall, demonstrating greater performance in detecting fraudulent job postings. To strengthen resilience and handle increasingly complex

VI. REFRENCES:

1. Augustine, Robin, VBIR-Based Assessment of Radiographic-Divergence Agent Attention in Prostate Melanoma Patients. Available at SSRN: https://ssrn.com/abstract=4752359 or http://dx.doi.org/10.2139/ssrn.4752359.

2. Detection of Fake Online Recruitment Using Machine Learning Techniques. Focused on ensemble techniques, this work uses supervised learning to differentiate fraudulent job postings from legitimate ones, showing high performance with Random Forest. Available on IEEE Xplore.

3. Rangaiah PKB, Histopathology-driven prostate cancer identification: A VBIR approach with CLAHE and GLCM insights. *Comput Biol Med.* 2024 Oct 1;182:109213. doi: 10.1016/j.compbiomed.2024.109213. Epub ahead of print. PMID: 39357133.

4. Fredrik Huss et al., Precision Diagnosis of Burn Injuries: Clinical Implications of Imaging and Predictive Modeling, 09 October 2024, PREPRINT (Version 1) available at Research Square https://doi.org/10.21203/rs.3.rs-5002889/v1.

5. Swaminathan, S. V., Surendiran, J., (2019). Design and Implementation of Kogge Stone adder using CMOS and GDI Design: VLSI Based. *International Journal of Engineering and Advanced Technology (IJEAT)*, 8(6S3).

6. Supervised Learning for Fake Job Detection. Focuses on supervised algorithms such as Support Vector Machines and Decision Trees for isolating fraudulent posts. It highlights the role of feature selection in enhancing model accuracy. Published in JES.

7. Robin Augustine, Enhancing Medical Image Reclamation for Chest Samples Using B-

Coefficients, DT-CWT and EPS Algorithm, in *IEEE Access*, vol. 11, pp. 113360-113375, 2023, doi: 10.1109/ACCESS.2023.3322205.

8. Using Multi-Layer Perceptron (MLP) for Detecting Fake Job Postings. MLP with layered architectures is discussed to classify job postings, optimizing for precision and recall to minimize false positives and negatives. Published in IRJET.

9. Naresh, E., Hemavathi, P., Padmavathi, S., Srinidhi, N. N., Pradeep Kumar, B. P., Karthik, V., & Mallik, S. (2024). Autonomous Garbage Accumulation Robot Using Internet of Things. *Journal of Machine and Computing*, 4(2), 431-440.

10. Srinidhi, N.N., Shiva Darshan, S.L. et al. Design of Cost Efficient VBIR Technique Using ICA and IVCA. *SN COMPUT. SCI.* 5, 560 (2024). doi: 10.1007/s42979-024-02936-9.

11. Employment Scam Detection Model Using Machine Learning. Implements logistic regression, Naive Bayes, and ensemble classifiers on datasets to build a robust prediction tool for identifying fake job posts. Published in IJERT.

12. Robin Augustine, Improving Liver Cancer Diagnosis: A Multifaceted Approach to Automated Liver Tumor Identification in Ultrasound Scans. Available at SSRN: https://ssrn.com/abstract=4646452 or http://dx.doi.org/10.2139/ssrn.4646452.

13. Machine Learning for Online Job Scam Detection. Explores advanced classifiers like Gradient Boosting and AdaBoost, along with feature engineering, to enhance scam detection models. Published in JES.

14. Data Preprocessing and Feature Engineering for Detecting Job Scams. Discusses text processing methods like tokenization, lemmatization, and TF-IDF vectorization to improve classifier performance. Published in IJERT.

15. Fake Job Recruitment Detection Using Machine Learning Approach. This research highlights the application of multiple classification models, including Random Forest, AdaBoost, and Naive Bayes, emphasizing ensemble classifiers' effectiveness in detecting fake job ads. Published in IJETT.

16. Pramod K B, Kumaraswamy H.V, Prathap C and M. Swamy, Design and analysis of UHF BJT feedback oscillator using linear and non-linear simulation, 2013 *International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA)*, Bangalore, India, 2013, pp. 1-6, doi: 10.1109/C2SPCA.2013.6749386.

17. Rangaiah, Pramod and Augustine, Robin, Enhanced Glaucoma Detection Using U-Net and U-Net+ Architectures Using Deep Learning Techniques. Available at SSRN: https://ssrn.com/abstract=4831407.

18. Ravikumar, J., Gauging Deep Learning Archetypal Effectiveness in Haematological Reclamation. *SN COMPUT. SCI.* 5, 963 (2024). doi: 10.1007/s42979-024-03322-1.

19. Darshan, S.L.S., Naresh, E. et al. Design of Chest Visual Based Image Reclamation Method Using Dual Tree Complex Wavelet Transform and Edge Preservation Smoothing Algorithm. *SN COMPUT. SCI.* 5, 352 (2024). doi: 10.1007/s42979-024-02742-3.

20. Manjunatha, M.B. (2018). Design and Development of ASL Recognition by Kinect Using Bag of Features. In: Reddy, M., Viswanath, K., K.M., S. (eds) *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*. Advances in Intelligent Systems and Computing, vol 628. Springer, Singapore. doi: 10.1007/978-981-10-5272-9_31.

21. Kumar, B. P. (2018). analysis of ASL recognition system for aligned RGB-D image. journal of advanced research in dynamical and control systems, (1).

22. Mallanna, S. D. (2018). High gain perfect matched inset fed rectangular microstrip patch antenna for 2.4 GHz frequency. J Adv Res Dyn Control Syst, 10(13-Special Issue), 46-52.

23. Manjunatha, M. (2016). Performance analysis of KNN, SVM and ANN techniques for gesture recognition system. Indian Journal of Science and Technology, 9(1), 1-8

24. Naresh, E., Hemavathi, P., Padmavathi, S., Srinidhi, N. N., Pradeep Kumar, B. P., Karthik, V., & Mallik, S. (2024). Autonomous Garbage Accumulation Robot Using Internet of Things. Journal of Machine and Computing, 4(2), 431-440

25. Kumar, B. P. (2019). Framework of ASL Silhouette Gesture Recognition System. In Blue Eyes Intelligence Engineering & Sciences Publication (Vol. 8, No. 6s, pp. 66-72)