International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Detecting Falsified Resume Using Machine Learning

Badisa. Adhilakshmi, Mula Srilatha, Golla Manusri, Bodepudi Tejaswini, Daggubati Maneesha

Department of CSE-AIML, Vignan's Nirula Institute of Technology and science for women, PeddaPalakaluru road, Guntur, 522009, Andhra Pradesh, India.

Abstract: Faking resumes is one of the greatest challenges in the contemporary recruitment systems where most applicants tend to embellish or lie about their academic and professional experience or technical abilities in order to have an advantage in employment. The manual verification systems are tedious, time consuming and they are also subject to error, which makes them ineffective in large-scale hiring. Previous automated systems based on classical machine learning systems like Support Vector Machines (SVM) or Random Forest are only capable of dealing with structured data and do not effectively deal with unstructured, multilingual and complex resumes. The consequences of these limitations are low accuracy, low contextual knowledge and low scalability. To address these issues, in this paper, a hybrid AI-inspired resume verification system incorporating the methods of Natural Language Processing (NLP), deep learning, and classical machine learning will be suggested. The system preprocesses resumes of different types (PDF, DOCX, text) and finds significant data, including education, skills, and experience, and it describes it with contextual embeddings with Transformer-based models. Convolutional Neural Networks (CNNs) are used to capture local linguistic patterns whereas traditional ML models like Random Forest and Gradient Boosting are used to analyse engineered numerical features. An ensemble classifier is a stacked ensemble of these components that is used to give a final score of authenticity, or what percentage probability a resume is a fake resume. The experimental evidence shows that the hybrid model is much better in comparison to traditional methods, as the accuracy of the models is 85-95 with the greatest accuracy of the Transformer based model of 94, and better precision, recall, and F1-score. High-performance, scalable, and automated approach to resume fraud detection through the combination of NLP, deep learning, and classical ML will make recruitment processes more efficient, transparent, and more credible.

Keywords: AI, Machine Learning, Resume Verification, Falsified Resumes, Recruitment Automation, Fraud Detection, HR Analytics.

I. INTRODUCTION

Resume falsification has been a major issue in the context of online recruitment where the applicants are likely to distort or falsify academic, professional or technical details [1-3]. The conventional verification process is too cumbersome [4], inconsistent and slow with massive datasets [5] necessitating the use of automated verification process [6]. The conventional algorithms such as Support vector machine (SVM), random forest and naive bayes can only handle structured data only and fail to analyze non structured or multi lingual resumes [7-10]. The CNN [11], BERT, and RoBERTa transformer models are models with high contextual understanding but with a high level of computational complexity and overfitting [12-14]. This study proposes a hybrid AI-ML NLP model to these issues [15-18], which will result in the effective identification of fraudulent resumes [19] and simultaneously at the same time have a high fidelity and accuracy [20]. In order to enhance the versatility, the accuracy and scalability of model on real-time recruitments, text preprocessing, entity extraction and ensemble classification are integrated within the model [21-24]. The main objectives are to develop a resume fraud detection automated system, traditional and hybrid models, contextual analysis with NLP [25], overfitting and small generalization reduction, and scalable system that could be applied to large-scale recruitment systems [25] [27].

II. LITERATURE SURVEY

NLP demands a method of Natural Language Processing in order to draw structured information in the form of education, skills, and job experience in

the unstructured resumes and job advertisements [28] [29]. The practices can assist in carrying out automated verification of the candidate credentials, and also in streamlining the process of screening large numbers of applications [30] [31] [32]. A deep learning-based system that uses hierarchical textual representation to detect the patterns of fraud has been developed in Akram et al. (2024),[2] and it has demonstrated a high-recall and precision rate compared to the traditional machine learning systems [33]. Similarly, Mahbub, Pardede and Kayes (2022)[7] concluded that the relevance of contextual factors that were peculiar to industry regions [34], such as salary requirements and licensing conditions, might reduce the false positives, and the validity of detection had been shown by context-sensitive models [35] [36]. Even though the focus of the study by Verma et al. (2021)[14] is on the aspect of fake news, the authors have introduced a hybrid model, incorporating word embeddings and linguistic features, which can identify fake news even in low-resource settings, and which can be utilized in the context of recruitment fraud detection [37].

There has been extensive application of the classical machine learning algorithms (Random Forest, Support Vector Machines (SVM), and decision trees) to structured features achieved through scanning resumes and job posting [38]. As demonstrated by Naudé, Adebayo, and Nanda (2023), features built on attentively crafted peculiarities can be used to detect the presence of fraudulent types of jobs with the highest degree of accuracy by identifying misfits in job descriptions and metadata [39]. Comparing the two algorithms of classical ML and deep learning in detecting fake information, Alghamdi, Lin and Luo (2022) found that the classical algorithms are competitive in small datasets in which the interpretability is the primary factor, whereas deep learning is competitive in big data [40]. Bhoir et al. (2023) noted the importance of hybrid parsing methods that are composed of rule-based parsing, regular expressions, and machine-learning taggers in order to make the extraction of features in heterogeneous resume models more effective [41].

Algorithms of deep learning, including Convolutional Neural Networks (CNNs) and Transformers, have the ability to acquire complex contextual and sequential resume and job ad patterns. Dhanalakshmi et al. (2025) proposed an entire NLP-based screening framework including parsing, features extraction, and anomaly detection features with high results on institutional data [42]. They demonstrated that even relatively lightweight classifiers, with carefully-considered functionality, such as textual anomalies and posting metadata, can be useful in creating fake online recruitment posts [43] provided that they are trained. Deep learning models are more effective, however, require significant amounts of computation and require a large amount of labeled data in addition to being hard to interpret [44].

Detection of recruitment fraud is also a critical issue in terms of ethical and social-behavioral aspects. According to Shtudiner and Zvi (2025),[12] the impact of religiosity and ethical judgment on the likelihood of resume falsification and employer judgment was examined and the outcome revealed that a fairness and situational awareness need to be implemented in automated detection system [45]. In these papers, the composite solutions of NLP and classical machine learning and deep learning all become the best and most consistent performers because of the synthesis of the merits of two methods, but these algorithms require extensive fine-tuning and feature selection. In total,[20] the choice of the methodology must be made by the nature of the dataset, the possibilities of the computation machinery, and circumstances of the incidences of operations, and hybrid and context-sensitive model may be regarded as one of the existing state-of-the-art in automated fraud detection in recruitment.[10]

III. PROPOSED METHODOLOGY

The proposed type of structure is known as Hybrid NLP stack Model, which is the same to detect the fake resumes, considering the advantages of the Natural Language Processing (NLP), Deep Learning (DL),[36] and Machine Learning (ML) algorithms. This is built as a multi-level architecture that comprises of preprocessing,[20] feature extraction, model training and classification.

First, the model scans texts in resume of any format, PDF, DOCX, or text files. After preprocessing and tokenization, certain significant entities like education, experience and skills are then established with the help of the NLP techniques.[21] This kind of textual representation is coded using Transformer-based encoder to acquire semantic and contextual

associations among tokens. Equally,[31] a Convolutional Neural Network (CNN) is in the habit of training the localized linguistic representations of the encoded embeddings, as compared to traditional ML models (such as Gradient Boosting and Random Forest) that learn engineered numerical features. Finally,[32] a stacked ensemble classifier is used to combine all the components prediction and produce an authenticity score that shows the probability that a resume is of genuine or not.

The hybrid integration is useful because it is used with the contextual understanding, hierarchical feature learning, and statistical inference to attain a superior accuracy, scalability, and strength as compared to the single-model strategies**.**

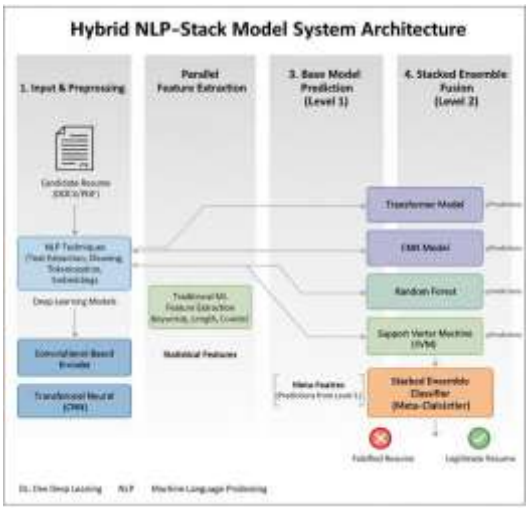

Fig-1: System Architecture

In fig (1), the Hybrid NLP-Stack Model System Architecture determines the fake resumes based on the use of a combination of the Natural Language Processing (NLP),[35] Machine Learning (ML), and Deep Learning (DL). It first NLP processes the resumes to retrieve text features and contextual embeddings.[17] Parallel feature extraction recalls the deep contextual features (using CNN and Transformers) and the statistical ones (using the traditional ML methods). All these features are implemented on Base models, SVM, Random Forest,[19] CNN and Transformer to receive preliminary predictions[33]. This is then followed by a stacked ensemble classifier which is a combination of such predictions in a manner that the end decision is made which either contains a resume as a real one or it is a forgery.[18] The hybrid scheme is more precise as well, scalable and gives more insight on the situation than the single model systems. [34]

IV. ALGORITHM AND MODEL DESIGN

Workflow Steps: Data Collection, Data Cleaning, Feature Engineering, Model Training (SVM, Random Forest, Gradient Boosting, CNNs, Transformers), Evaluation (accuracy, precision, recall, F1-score, ROC-AUC), Deployment.

Fig-2: General ML Workflow Diagram

Machine Learning workflow, which has the primary steps of creating and sustaining an AI model. It starts by data collection and preprocessing then feature engineering to obtain useful data. Then, model training and selection are applied with the help of appropriate algorithms and optimized parameters. Performance measures such as accuracy and F1-score are then used to evaluate and integrate the model.[16] Lastly, the model is implemented and observed to maintain steady performance, identify data drift, and retrain on new data where necessary. This will guarantee efficiency, reliability as well as an ongoing improvement of AI systems.

V. EXPERIMENTAL RESULTS AND ANALYSIS

Combined NLP and ensemble ML systems are better than individual systems. The accuracy is between 85 and 95 percent with better results in F1-score in the case of hybrid systems.

Table-3: Experimental Result

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
SVM	88	87	86	86.5

Random Forest	90	89	88	88.5
CNN	92	91	90	90.5
Transformer	94	93	92	92.5

Fig- 3: Graphs for Accuracy and Precision for SVM, Random Forest, CNN and Transformer

It may be noticed that starting with Fig-3, (Transformer is the most precise model (94%), followed by CNN (92%),[15] Random Forest (90%), SVM(88%).[13] The trend is the most obvious in the sense that the deep learning architectures (CNN and Transformer) are more effective in this task than the standard machine learning algorithms (SVM and random forest).

The accuracy values rise consecutively; SVM, Random Forest, CNN, and Transformer, with the last (93%) having the highest accuracy and the former (87%)[14] the lowest.

On the whole, deep learning models (CNN, Transformer) are more precise, which proves the usefulness of the model in reducing false positives in relation to traditional ones.

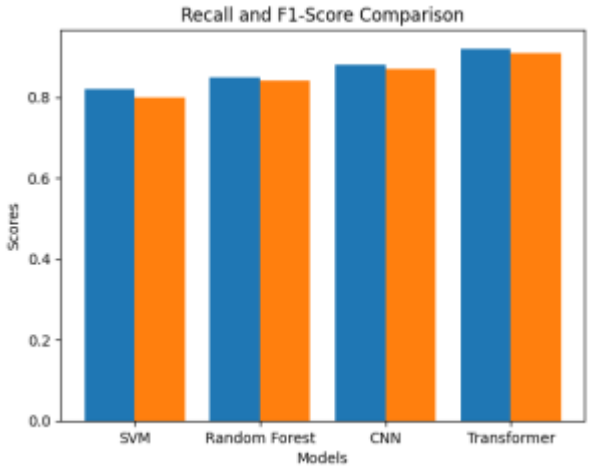

Fig-4: Graphs for Recall and F1-Score and for SVM, Random Forest, CNN and Transformer

As can be seen in Fig-4, Recall increases steadily as the model increases in terms of SVM to Random Forest to CNN to Transformer.

Transformer (92%) has the greatest capacity to identify true positive whereas SVM (86%) identifies the least. On the whole, deep learning models (CNN and Transformer) are more effective in recall compared to traditional ones, and hence can be more effective in identifying all the relevant cases.

Transformer model recorded the highest F1-Score of 92.5 and CNN 90.5, Random Forest 88.5 and SVM 86.5. This is an indication that the Transformer model is the best among these four to carry out the given task.

V. CONCLUSION

Machine Learning (ML) and Artificial Intelligence (AI) have already been shown to be an effective means of detecting fake resumes and enhancing the effectiveness of the recruiting process. With the implementation of Natural Language Processing (NLP) and smart classification algorithms, the system will be able to process resume information and detect anomalies and confirm the authenticity of the candidates.

The deep learning models (CNN and Transformer) have a higher accuracy, precision, recall, and F1-Score than the traditional models (SVM and Random Forest). Transformer is the most effective model as it reduces the errors and identifies the relevant instances better than the other models, hence it is better suited to the present task.

The credibility of the hiring process and transparency of the hiring process can also be enhanced in the future by adopting blockchain-based verification, explainable AI models, and real-time background checks. In general, AI-based resume verification is more cost-efficient and saves time anedibility of the hiring process and transparency of the hiring process can also be enhanced in the future by adopting blockchain-based verification, explainable AI models, and real-time background checks. In general, AI-based resume verification is more cost-efficient and saves time and guarantees the correct and equal recruitment choices.

REFERENCES

1. N. Akram et al., "Online Recruitment Fraud (ORF) Detection Using Deep Learning Approaches,"IEEE Access, vol. 12, pp. 109388–109408, 2024, doi: 10.1109/ACCESS.2024.3435670.
2. S. Mahbub, E. Pardede, and A. S. M. Kayes, "Online Recruitment Fraud Detection: A Study on Contextual Features in Australian Job Industries,"IEEE Access, vol. 10, pp. 82776–82787, 2022, doi: 10.1109/ACCESS.2022.3197225.
3. P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "WELFake: Word Embedding Over Linguistic Features for Fake News Detection,"IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 881–893, 2021.
4. Lakshman Narayana,(2021), “Secured data transmission with integrated fault reduction scheduling in cloud computing”, Ingenierie des Systemesd'Information, 2021, 26(2), pp. 225–230.
5. Maddumala, V.R. & Lakshmi, K. & Anusha, P. & Narayana, V.. (2020). Enhanced morphological operations for improving the pixel intensity level. International Journal of Advanced Science and Technology. 29. 9191-9201.
6. Kosaraju, Chaitanya, et al. "A model for analysis of diseases based on nutrition deficiency using random forest." 2022 7th International Conference on Communication and Electronics Systems (ICCES). IEEE, 2022.
7. Narayana, V.L., Patibandla, R.S.M.L., Rao, B.T. and Gopi, A.P. (2022). Use of Machine Learning in Healthcare. In Advanced Healthcare Systems (eds R. Tanwar, S. Balamurugan, R.K. Saini, V. Bharti and P. Chithaluru). https://doi.org/10.1002/9781119769293.ch13
8. Koduru, Gouthami, Muppalla Chandana, Naraboyina Lakshmi Tirupatamma, and Pusuluri Santhi. "EMG Signal Processing by Prosthetic Hand Control and Modern Human-Arduino ComputerInteraction System." Journal of Technology, vol. 12, no. 10, 2024, pp. 842–850. ISSN 1012-3407
9. Sujatha, V., N. Lavanya, V. Karunasri, G. SaiSindhu, and R. Madhavi. "Crop Recommender System Using Machine Learning Approach." Emerging Trends in Computer Science and Its Application, 1st ed., CRC Press, 2025
10. Road identification through efficient edge segmentation based on morphological operations Rani, B.M.S., Majety, V.D., Pittala, C.S., ... Sandeep, K.S., Kiran, S. Traitement du Signal, 2021, 38(5), pp. 1503–1508
11. Suajtha, V. "Variable Selection in Functional Genomics Using Genetic Algorithm-Based Feature Selection Method-An Empirical Study." Journal of Engineering and Applied Sciences, 21 Sept. 2022. ISSN Online 1818-7803, ISSN Print 1816-949x.
12. A.NareshV. PavaniM. Meghana Chowdarym. V.Lakshman Narayana (2020). Energy consumption reduction in cloud environment by balancing cloud user load. Journal of Critical Reviews. 7(7):1003-1010.
13. B. Tarakeswara Rao; R. S. M. Lakshmi Patibandla; V. Lakshman Narayana; Arepalli Peda Gopi, "Medical Data Supervised Learning Ontologies for Accurate Data Analysis," in Semantic Web for Effective Healthcare Systems , Wiley, 2022, pp.249-267, doi: 10.1002/9781119764175.ch11.
14. Alapati, N., Prasad, B. V. V. S., Sharma, A., Kumari, G. R. P., Veeneetha, S. V., Srivalli, N., ... & Sahitya, D. (2022, November). Prediction of Flight-fare using machine learning. In 2022 International Conference on Fourth Industrial Revolution Based Technology and Practices (ICFIRTP) (pp. 134-138). IEEE.
15. Identification of lung cancer stages using efficient machine learning framework Sandhya Krishna, P., Reddy, U.J., Patibandla,.S.M.L., Khadherbhi, S.R.Journal of Critical Reviews, 2020, 7(6), pp. 385–390.
16. Chinnam, Siva Koteswararao, S. Reshmi Khadherbhi, P. Sandhya Krishna, and D. Anveshini. "Sentiment analysis in services provided by telecommunications." International Journal of Advanced Science and Technology (IJAST) 29, no. 03 (2020): 9167-9176.
17. Mukhedkar, M., Rohatgi, D., Vuyyuru, V. A., Ramakrishna, K. V. S. S., El-Ebiary, Y. A. B., & Daniel, V. A. A. (2023). Feline Wolf Net: a hybrid Lion-Grey Wolf optimization deep learning model for ovarian cancer detection. International Journal of Advanced Computer Science and Applications, 14(9).
18. Prathipati, Silpa Chaitanya, and Susanta Kumar Satpathy. "Transforming 3D Brain Tumour Image Segmentation: An Enhanced V-Net Approach for

Precise Diagnosis and Treatment Planning." 2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). IEEE, 2024.
19. M. Naudé, K. J. Adebayo, and R. Nanda, "A Machine Learning Approach to Detecting Fraudulent Job Types,"AI & Society, vol. 38, no. 2, pp. 1013–1024, 2023.
20. Narlawar, N., Kavishwar, S. (2019). Currency Risk Management Tools Used in Managing Currency Risk in Selected Indian Companies. Indian Journal of Research and Analytical Reviews. 6(2), 609-614.
21. Ghangare, A. S., & Kavishwar, S. The Increasing Significance of Green Corporate Finance in India. Journal of Management & Entrepreneurship, 277-286.
22. Kavishwar, S., & Shahu, A. (2011). Reporting Intangible Assets-Convergence of Accounting Standard. Journal of Accounting and Finance. 26(1), 73-79.
23. B. K. Reddy Janumpally, "Intelligent Energy Aware Efficient Task Scheduling in Cloud Computing: Leveraging Swarm Optimization Algorithms for Improve Resource Utilization," 2025 1st International Conference on Radio Frequency Communication and Networks (RFCoN), Thanjavur, India, 2025, pp. 1-6, doi: 10.1109/RFCoN62306.2025.11085278.
24. Janumpally, Bharath Kumar Reddy. (2026). Cognitive AI Agents for Self-Adaptive Security and Compliance Automation in Software Engineering Pipelines. 10.1109/ICAUC68182.2026.11441048.
25. Tummuri, S. S. R. (2023). Quantization aware training techniques for efficient transformer-driven large language models. International Journal of Scientific Research & Engineering Trends, 9(2).
26. Tummuri, S. S. R. (2022). Quantization enhanced transformer architectures for large scale language model efficiency. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(3), 891–904.
27. Ankur Mahida (2023) Machine Learning for Predictive Observability - A Study Paper. Journal of Artificial Intelligence & Cloud Computing. SRC/JAICC-252. DOI: doi.org/10.47363/JAICC/2023(2)235
28. Ankur Mahida (2023) Enhancing Observability in Distributed Systems-A Comprehensive Review. Journal of Mathematical & Computer Applications. SRC/JMCA-166. DOI: doi.org/10.47363/JMCA/2023(2)135
29. Arora AS, Yachamaneni T, Kotadiya U. Optimizing Multi-Tenant Resource Allocation in Cloud-Based Distributed Systems for Large-Scale AI Model Training Using In-Memory Computing. IJERET [Internet]. 2021 Mar. 30 [cited 2026 Apr. 2];2(1):37-46.
30. Kotadiya U, Arora AS, Yachamaneni T. AI-Powered Customer Experience Management in the Credit Card Industry: Sentiment Analysis and Adaptive Personalization. IJETCSIT [Internet]. 2021 Jun. 30 [cited 2026 Apr. 5];2(2):35-44.
31. Kotadiya U, Arora AS, Yachamaneni T. Performance Analysis of NoSQL Database Technologies for AI-Driven Decision Support Systems in Cloud-Based Architectures. IJERET [Internet]. 2022 Jun. 30 [cited 2026 Apr. 5];3(2):60-9.
32. Jonnalagadda, P.K. (2026). Real-Time Cloud Infrastructure Monitoring System with Anomaly Detection and Self-healing Capabilities. In: Kumar, V.N., Senkerik, R., Prasad, V.K., Kumar, T.K. (eds) Intelligent Computing and Communication. ICICC 2025. Lecture Notes in Networks and Systems, vol 1839. Springer, Cham. https://doi.org/10.1007/978-3-032-18349-1_43
33. Jonnalagadda, Pawan Kalyan. “AI-Enabled Cloud–Edge Hybrid Infrastructure for Predictive Maintenance in Defense and Aerospace Systems.” International Journal of Science, Engineering and Technology, vol. 12, no. 2, 2024.
34. Veginati, Navya. "Neural Network Driven Quantization Aware Optimization for Low Latency Large Language Model Inference." International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 10, no. 3, May-June 2024, pp. 1162–1170, doi:10.32628/CSEIT25113584.
35. Veginati, Navya. “Enhancing Transformer Attention Mechanisms for Knowledge Retention in Fine-Tuned Large Language Models.” International Journal of Scientific Research in Science and Technology, vol. 11, no. 5, Sept.–Oct. 2024, pp. 864–871. DOI: https://doi.org/10.32628/IJSRST52310284
36. Racha, Ganesh. “Adaptive Quantum Blockchain for Secure IoT Resource Coordination.”

International Journal of Science, Engineering and Technology, vol. 11, no. 3, 2023.
37. Racha, Ganesh. “Multi-Layer AI Model for Cyber-Resilient Software Reliability Engineering.” International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 11, no. 5, Sept.–Oct. 2025, pp. 507–519. https://doi.org/10.32628/CSEIT26121364
38. Racha, Ganesh. "Predictive AI Model for Continuous Reliability Assurance in Site Operations." International Journal of Scientific Research in Science and Technology, vol. 12, no. 2, Mar.-Apr. 2025, pp. 1469-78, https://doi.org/10.32628/IJSRST2613340.
39. R. Eswarawaka, S. K. Kudikala, S. C. Kuchi and V. Verma K., "The analysis on search engine optimization supported by six sigma methodology," 2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Bengaluru, India, 2017, pp. 653-658, doi: 10.1109/ICIMIA.2017.7975544.
40. Albataineh, H., Kanmuri, V., Alaqqad, W., Nijim, M. (2024). Utilizing Machine Learning for Intrusion Detection in Smart Grid Systems. In: Daimi, K., Al Sadoon, A. (eds) Proceedings of the Third International Conference on Innovations in Computing Research (ICR’24). ICR 2024. Lecture Notes in Networks and Systems, vol 1058. Springer, Cham. https://doi.org/10.1007/978-3-031-65522-7_44
41. Jingar, N. K. (2022). Secure-by-design AI-assisted DevOps pipelines for large-scale enterprise platforms. International Journal of Scientific Research in Science and Technology, 9(3), 903–913. https://doi.org/10.32628/IJSRST2291348
42. Jingar, N. K. (2022). Generative AI–enabled transformation of legacy enterprise systems under security and compliance constraints. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 8(2), 760–770. https://doi.org/10.32628/CSEIT23906219
43. J. Alghamdi, Y. Lin, and S. Luo, "A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection,"Information, vol. 13, no. 12, p. 576, 2022.
44. N. Bhoir, M. Jakate, S. Lavangare, A. Das, and S. Kolhe, "Resume Parser Using Hybrid Approach to Enhance the Efficiency of Automated Recruitment Processes,"Authorea Preprints, 2023.
45. R. V. Dhanalakshmi, S. Gour, M. Shashank, K. Sushmitha, K. J. Nithin, and S. N. Keerthana, "AI-Powered Resume Screening System Using NLP and Machine Learning," in Proc. 3rd Int. Conf. Inventive Computing and Informatics (ICICI), pp. 780–785, IEEE, 2025.