Streamlining Job Search and Detecting Fraudulent Job Listing

1st Kunal Mahendra Kachare
Department of Information Technology
Pune Institute of Computer Technology
Pune, India

2nd Akshay Ashok Koganur
Department of Information Technology
Pune Institute of Computer Technology
Pune, India

3rd Avinash Laxman Maharnavar
Department of Information Technology
Pune Institute of Computer Technology
Pune, India

4th Mihir Tushar Rajpathak
Department of Information Technology
Pune Institute of Computer Technology
Pune, India

5th Mrs. Rachana.A.Karnavat
Department of Information Technology
Pune Institute of Computer Technology
Pune, India

***Abstract*—In today's job market to explore how various contextual factors influence the accuracy of automatically detecting online recruitment fraud, this study seeks to highlight the significance of tailoring such fraud detection methods to local contexts. To achieve these goals, the study initially creates a dataset using a local and semi-structured advertising platform. This dataset is then utilized to train a machine learning model, which considers a range of content-based and contextual features. The study's findings indicate that the incorporation of contextual features enhances the performance of automated online recruitment fraud detection models. The practical implications of this research are twofold. First, the contextual feature generation engine can be applied to diverse data-sets with minimal localization efforts. Second, these machine learning models can be integrated into online job recruitment platforms to identify and prevent online recruitment fraud.**

***Index Terms*—Machine learning, web crawling, natural langupage processing (NLP).**

I. Introduction

In the contemporary digital landscape, the way we seek employment has undergone a profound shift. The emergence of online job platforms has reshaped the job-seeking experience, offering countless opportunities, but it has also brought to the fore a concerning challenge - the prevalence of potentially fraudulent job listings. These deceptive postings pose substantial risks to job seekers, making it crucial to address this issue with innovative solutions.

Our ambitious project has embarked on a two-fold mission: to bolster user safety and streamline the job search process. In an era where efficiency and security are paramount, we aim to empower job seekers with a safer and more efficient experience. Our automated system is designed to identify and flag suspicious job listings, protecting users from scams, fraud, and misleading opportunities. This endeavor not only safeguards job seekers but also upholds the credibility and trustworthiness of digital job platforms. At the core of our system is its ability to meticulously analyze job listings. It examines various elements, including keywords, descriptions, and company details, to distinguish between legitimate opportunities and those that raise red flags. What makes our approach stand out is its reliance on data-driven analysis, supported by state-of-the-art machine learning models. This dynamic aspect ensures that our system adapts and improves over time, staying one step ahead of those with malicious intentions.

In the ever-evolving landscape of online job listings, it is essential to have a system that can adapt and grow in tandem with the changing tactics employed by fraudulent actors. This is where machine learning enters the picture. Our system employs advanced machine learning models to continually refine its ability to detect suspicious listings. This project recognizes that an effective solution to this challenge requires a combination of cutting-edge technology and a commitment to user welfare. By employing web crawling techniques, our system scours online job platforms to identify listings that exhibit suspicious characteristics. These suspicious traits can range from unusual keywords and vague descriptions to discrepancies in company details. Crucially, we don't stop at mere identification; we take the extra step of flagging these listings, allowing job seekers to approach them with caution. This proactive approach not only shields users from potential harm but also empowers them to make more informed decisions about the job opportunities they pursue.

Our commitment to this project is unwavering, fueled by the vision of creating an environment where job seekers can navigate the digital job market with absolute confidence. We envision a future where fraudulent job listings are a relic of the past, and job seekers can focus on finding the right opportunities with ease, free from the shadows of scams and deceit. Our journey to transform the job-seeking landscape is an exciting one, and it reflects our dedication to making the digital age safer, more efficient, and full of opportunities for all.

II. Literature Survey

According to the several studies, researchers have made significant strides by developing innovative solutions leveraging a range of technologies. One noteworthy study, focusing on the

Australian job industry [1], delves into the impact of automatic detection accuracy for online fake job listings within the local job market. This study's core foundation lies in the creation of a labeled dataset, meticulously curated from local online job postings, which serves as the training material for detecting fraudulent job listings. What sets this research apart is its emphasis on using both contextual and content-based features. The study illuminates an effective approach for automatically extracting these features from various online job listing platforms. By doing so, it equips fraud detection algorithms with a holistic understanding of the job postings, making them more adept at distinguishing genuine opportunities from fraudulent ones.

The research with dataset EMSCAD, [2] a data selection process involves gathering data with a focus on different types, including real, identity, corporate, marketing, and corporate identity data. Data annotation is performed to categorize and label the collected data. Feature selection is then applied, with options ranging from bag of words models, empirical rulesets, word embeddings, to more advanced techniques like transformers. Finally, for analysis, a variety of machine learning algorithms are considered, such as Logistic Regression (LR), Stochastic Gradient Descent (SGD), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forests (RF), AdaBoost (AB), and Classification and Regression Trees (CART) [3]. This comprehensive approach allows for thorough data exploration and analysis within the EMSCAD framework. The FNC-1 dataset [4] encompassing four categories (agree, discuss, disagree, unrelated), was employed for fake news detection. Feature extraction methods applied included N-grams, HashingTF-IDF for text-to-numerical conversion, and count vectorization to create a token count matrix. To enhance performance, a stacked ensemble model was employed, achieving an impressive F1 score of 92.45%, signifying a substantial 9.35% improvement over prior techniques in the domain of fake news detection. This advancement underscores the efficacy of the chosen dataset, feature extraction techniques, and the power of ensemble modeling in accurately identifying fake news.

[5] A diverse range of datasets, including PolitiFact, GossipCop, Twitter's Election Integrity, FakeNewsNet, Buzzfeed, Weibo, Liar, PHEME, ISOT, and the COVID-19 Pandemic (2020), are used for the task at hand. Feature extraction is accomplished using Principal Component Analysis (PCA), while a firefly-optimized algorithm is employed for feature selection. The considered features span linguistic characteristics such as word, character, and sentence averages, semantic attributes involving Latent Dirichlet Allocation (LDA) topic and emotion distributions, and network-related properties based on nodes, edges, and degree distributions. A suite of machine learning algorithms is applied, [6] including Multi-Support Vector Machine (SVM) Classification with high accuracy in multiple datasets, Bagging SVM, Boosting SVM, Rotation Forest Classifier, and a Stacking approach. This comprehensive methodology demonstrates a sophisticated and multifaceted approach to leveraging feature engineering and diverse datasets for the purpose of classification and analysis.

[7] In the context of fake news detection using the ”my-COVID19” dataset, a two-stage algorithmic approach is employed. In the first stage, Genetic Programming (GP) is utilized to generate mathematical expressions based on a set of 25 features, including metrics like average word length and unique word count. GP uses a tree-based representation to evolve these expressions, creating an initial mathematical model for fake news detection. Subsequently, in the second stage, Adaptive Differential Evolution (ADE), a variant of the differential evolution algorithm, is employed to fine-tune and optimize the mathematical model generated by GP. This combination of GP and ADE aims to enhance the accuracy and reliability of fake news detection by evolving and optimizing mathematical equations to capture relationships within the dataset and improve overall performance.

The comprehensive survey [8], [9] on web scraping techniques and applications delved into a wide spectrum of methods, spanning from time-honored traditional approaches to cutting-edge modern techniques, all rooted in the realms of web crawling and web page parsing. This in-depth exploration sheds light on the evolution and diversity of tools and strategies employed in the dynamic field of web scraping, providing valuable insights into the intricacies and real-world applications of this pivotal technology.

Web scraping techniques can be categorized as manual or automated methods. Manual techniques involve copying and pasting data, while automated methods include HTML and DOM parsing, as well as the use of web scraping software. When scraping data for machine learning, the process typically involves identifying relevant data sources, selecting appropriate scraping tools, writing scalable code, cleaning the data, storing it in a suitable format, optional data labeling, and using the data to train machine learning models. It's important to adhere to legal and ethical considerations while web scraping, such as respecting website terms of service and data privacy laws [10]. Traditional web crawling stores all web pages in databases, leading to unnecessary storage of web pages and increased time for constructing sentiment dictionaries. [11] To tackle these issues, a sentiment-aware web crawling approach is introduced, featuring two hash-based methods: hash join and bucket-sorted hash join. The latter, a novel method, is particularly efficient. Experimental results demonstrate that the bucket-sorted hash join significantly outperforms existing web crawling methods, reducing both running time and storage space. This method accomplishes sentiment-aware tasks in 0.016 seconds per web page and saves 59% of the database space when compared to conventional web crawling techniques.

The increasing demand for positions at financially stable and promising technical companies and startups, highlighting the challenge job seekers face in navigating vast and often irrelevant job listings. To address this, a data-centric system is proposed, focusing on data quality over quantity. [12] The system employs custom data collection techniques, including web crawling and web scraping, to gather relevant data for job seekers. Different types of web crawlers, such as job listing, ontology-based, HTML, and API crawlers, are used to automate the process and reduce human effort and errors. Additionally, the text introduces recommendation systems as a means of filtering and segregating information for users, including collaborative filtering and content-based filtering techniques, to help users find job recommendations tailored to their preferences and interests.

III. Proposed Methodology

The methodology proposed by us gives an automated system that parses through the jobs and checks whether those jobs are fake or not. Users are shown authentic and unique jobs saving their time searching for jobs and applying as well as saving them from fraudulent activities. The web extension scrapes the current search pages of a job portal. It identifies and extracts all the job links available on that page. The job links are temporarily stored in a storage system for further processing. Each link is crawled and the job description of that job is checked and the models trained on Machine Learning and Deep Learning techniques will classify each job description as authentic or fraudulent based on the patterns and features they've learned during training. The authentic job links detected by the models are presented to the user as an output. These are the job postings that have been classified as genuine and trustworthy.

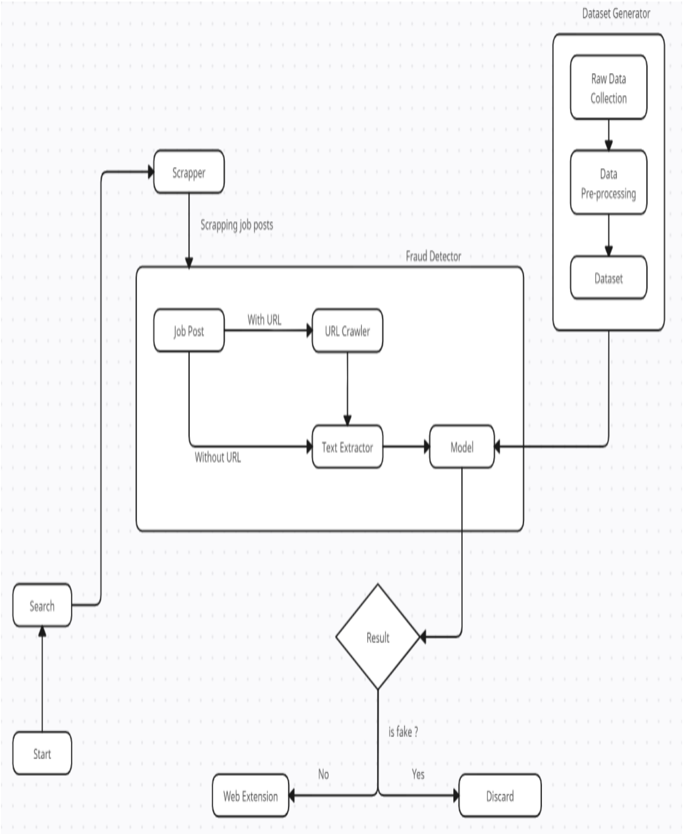

Fig. 1. Data Flow Diagram : Fraud job Detection

1. Dataset Generation : A dataset of the job descriptions for detecting fake jobs will be created from multiple tech job URLs and descriptions. The dataset consists of authentic URLs of career portals of companies, job descriptions of authentic tech jobs collected from the authentic career portal of authentic companies as well as fake job descriptions from fake job postings and datasets on popular platforms like Kaggle, UCI, etc.

2. Detecting important keywords and semantics of job descriptions : Natural Language Processing (NLP) is used for analyzing and processing data. This includes tasks like text classification, sentiment analysis, and named entity recognition. NLTK library enriches text processing. Techniques like TF-IDF, and Bag of Words will be used for feature extraction in NLP. TF-IDF is used to determine the importance of a word within a document relative to a collection of documents. Bag of Words stores the words in the bag and keeps the count of each word.

3. Model Building : Machine Learning algorithms like Logistic Regression, Random Forest as well as Deep Learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) will be used to train the model on the dataset for detecting the fake job postings. Logistic regression provides interpretable results, making it easier to understand which features contribute to the classification decision. Random Forest is robust to overfitting and can handle noisy or unstructured textual data giving better results. CNNs can be applied to process and extract features from text, such as job descriptions, by treating them as images in a 1D convolutional manner. RNNs can model sequential data in job postings, such as the order of sentences or phrases. LSTMs can capture complex relationships and dependencies within job descriptions, which may be important for identifying fake job postings.

4. Web Scraper : Web Scraper will scrap the current page of the job application portal. It will detect the structure of posts and the links that are present in the posts. Web Scrapper will also detect the content in a job post like description, requirements, and company name and that content will be sent as input to the model to detect fake jobs and authentic jobs. HTML parsing, DOM parsing, and Semantic annotation recognition are types of Web scraping techniques that can be used to scrap websites effectively. HTML parsing libraries navigate the HTML document tree and locate specific elements, such as job postings, within the page's structure. DOM parsing focuses on the structured representation of the web page in memory. Semantic annotation recognition helps to understand the semantic meaning of the content on a webpage.

5. AWS : AWS offers a vast array of cloud services, including computing power, storage, databases, and machine learning. AWS provides flexibility, scalability, and a pay-as-you-go pricing model. AWS S3(Simple Storage Service) is an object storage service for storing and retrieving data, including images, videos, and documents. AWS S3 will temporarily store job links which will be later crawled.AWS EC2(Elastic Compute Cloud) is a web service provided by Amazon Web

Services (AWS) that can run virtual machines (instances) in the cloud. EC2 provides resizable computing capacity in the cloud, making it a fundamental building block for many cloud-based applications.

IV. System Architecture

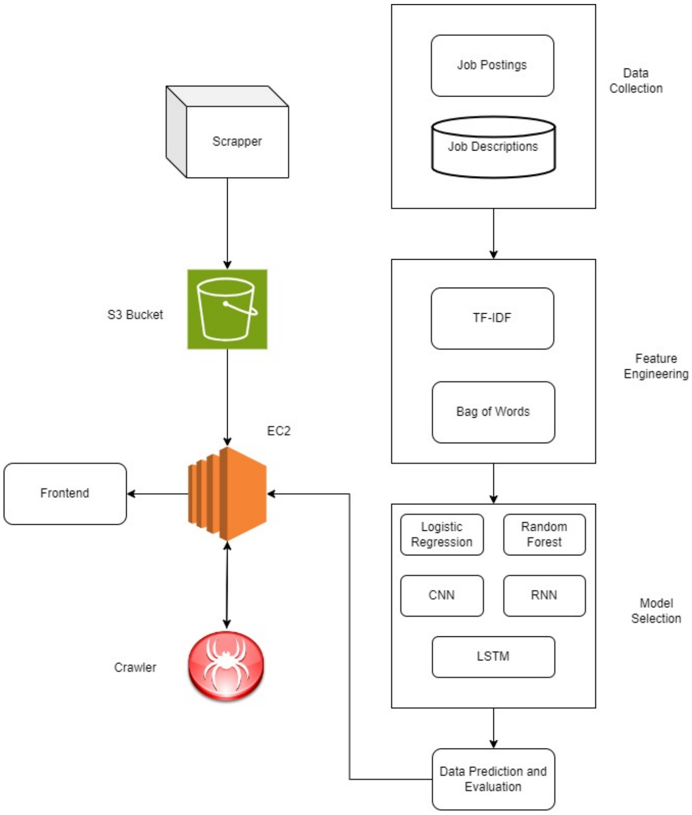

Fig. 2. System Architecture : Fraud Job Detection

V. Conclusion and Future Scope

In conclusion, as these systems continue to evolve, they hold the promise of reshaping how individuals find employment and organizations identify talent. This survey provides a comprehensive understanding of their current state and future prospects, serving as a valuable resource for all stakeholders interested in enhancing the efficiency, security, and fairness of the job search process. The integration of these systems stands to revolutionize job hunting, making it more convenient, secure, and inclusive for everyone involved.

The automated job detection platform can be further extended to show only those jobs that are relevant to the experience and level of the user. This will save a lot of time for employees who are in search of jobs and want to join a company on an urgent basis. Currently, the platform works on jobs from the Software Engineering field. The platform can be further extended for other professions by collecting job descriptions of their field and detecting fraudulent jobs in those fields. Increasing the performance of the models developed can be done in future studies. Multiple factors can be considered to detect fraudulent jobs in the future like images in job links, company descriptions, and role-based and experienced-based requirements in the job descriptions.

References

[1] S. Mahbub, E. Pardede & A. S. M. Kayes, ”Online Recruitment Fraud Detection: A Study on Contextual Features in Australian Job Industries,” in IEEE Access, vol. 10, pp. 82776-82787, 2022, doi: 10.1109/ACCESS.2022.3197225.

[2] Naudé, M., Adebayo, K.J. & Nanda, R. A machine learning approach to detecting fraudulent job types. AI and Soc 38, 1013–1024 (2023). https://doi.org/10.1007/s00146-022-01469-0.

[3] Bandyopadhyay, Samir & Dutta, Shawni. (2020). Fake Job Recruitment Detection Using Machine Learning Approach. International Journal of Engineering Trends and Technology. 68. 10.14445/22315381/IJETT-V68I4P209S.

[4] M. Park & S. Chai, ”Constructing a User-Centered Fake News Detection Model by Using Classification Algorithms in Machine Learning Techniques,” in IEEE Access, vol. 11, pp. 71517-71527, 2023, doi: 10.1109/ACCESS.2023.3294613.

[5] A. Altheneyan & A. Alhadlaq, ”Big Data ML-Based Fake News Detection Using Distributed Learning,” in IEEE Access, vol. 11, pp. 29447-29463, 2023, doi: 10.1109/ACCESS.2023.3260763.

[6] Ravish, R. Katarya, D. Dahiya & S. Checker, ”Fake News Detection System Using Featured-Based Optimized MSVM Classification,” in IEEE Access, vol. 10, pp. 113184-113199, 2022, doi: 10.1109/ACCESS.2022.3216892.

[7] J. T. H. Kong, W. K. Wong, F. H. Juwono & C. Apriono, ”Generating Fake News Detection Model Using A Two-Stage Evolutionary Approach,” in IEEE Access, vol. 11, pp. 85067-85085, 2023, doi: 10.1109/ACCESS.2023.3303321.

[8] L. Ying, H. Yu, J. Wang, Y. Ji and S. Qian, ”Fake News Detection via Multi-Modal Topic Memory Network,” in IEEE Access, vol. 9, pp. 132818-132829, 2021, doi: 10.1109/ACCESS.2021.3113981.

[9] Lotfi, Chaimaa & Srinivasan, Swetha & Ertz, Myriam & Latrous, Imen. (2021). Web Scraping Techniques and Applications: A Literature Review. 10.52458/978-93-91842-08-6-38.

[10] S. D. S. Sirisuriya, ”Importance of Web Scraping as a Data Source for Machine Learning Algorithms - Review,” 2023 IEEE 17th International Conference on Industrial and Information Systems (ICIIS), Peradeniya, Sri Lanka, 2023, pp. 134-139, doi: 10.1109/ICIIS58898.2023.10253502.

[11] B. -W. On, J. -Y. Jo, H. Shin, J. Gim, G. S. Choi and S. -M. Jung, ”Efficient Sentiment-Aware Web Crawling Methods for Constructing Sentiment Dictionary,” in IEEE Access, vol. 9, pp. 161208-161223, 2021, doi: 10.1109/ACCESS.2021.3129187.

[12] Kumar N, Gupta M, Sharma D, Ofori I. Technical Job Recommendation System Using APIs and Web Crawling. Comput Intell Neurosci. 2022 Jun 21;2022:7797548. doi: 10.1155/2022/7797548. PMID: 35774438; PMCID: PMC9239795.

[13] J. Bergman & O. B. Popov, ”Exploring Dark Web Crawlers: A Systematic Literature Review of Dark Web Crawlers and Their Implementation,” in IEEE Access, vol. 11, pp. 35914-35933, 2023, doi: 10.1109/ACCESS.2023.3255165.

[14] W. Shahid, Y. Li, D. Staples, G. Amin, S. Hakak and A. Ghorbani, ”Are You a Cyborg, Bot or Human?—A Survey on Detecting Fake News Spreaders,” in IEEE Access, vol. 10, pp. 27069-27083, 2022, doi: 10.1109/ACCESS.2022.3157724.

[15] M. Tajrian, A. Rahman, M. A. Kabir & M. R. Islam, ”A Review of Methodologies for Fake News Analysis,” in IEEE Access, vol. 11, pp. 73879-73893, 2023, doi: 10.1109/ACCESS.2023.3294989.

[16] C. Prashanth, D. Chandrasekaran, B. Pandian, K. Duraipandian, T. Chen and M. Sathiyanarayanan, ”Reveal: Online Fake Job Advert Detection Application using Machine Learning,” 2022 IEEE Delhi Section Conference (DELCON), New Delhi, India, 2022, pp. 1-6, doi: 10.1109/DELCON54057.2022.9752784.

[17] B. Pandey, T. Kala, N. Bhoj, H. Gohel, A. Kumar and P. Sivaram, ”Effective Identification of Spam Jobs Postings Using Employer Defined Linguistic Feature,” 2022 1st International Conference on AI in Cybersecurity (ICAIC), Victoria, TX, USA, 2022, pp. 1-6, doi: 10.1109/ICAIC53980.2022.9897059.

[18] D. Rohera et al., ”A Taxonomy of Fake News Classification Techniques: Survey and Implementation Aspects,” in IEEE Access, vol. 10, pp. 30367-30394, 2022, doi: 10.1109/ACCESS.2022.3159651.

[19] A. Mandadi, S. Boppana, V. Ravella & R. Kavitha, ”Phishing Website Detection Using Machine Learning,” 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824801.

[20] S. A. Alshalif et al., ”Alternative Relative Discrimination Criterion Feature Ranking Technique for Text Classification,” in IEEE Access, vol. 11, pp. 71739-71755, 2023, doi: 10.1109/ACCESS.2023.3294563.

[21] T. Bhatia and J. Meena, ”Detection of Fake Online Recruitment Using Machine Learning Techniques,” 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 300-304, doi: 10.1109/ICAC3N56670.2022.10074276.

[22] Sultana Umme Habiba, Md Khairul Islam and Farzana Tasnim, ”A comparative study on fake job post prediction using different data mining techniques”, 2021 2nd International Conference on Robotics Electrical and Signal Processing Techniques (ICREST), pp. 543-546, 2021.

[23] Ibrahim M. Nasser, Amjad H. Alzaanin and Ashraf Yunis Maghari, ”Online Recruitment Fraud Detection using ANN”, Palestinian International Conference on Information and Communication Technology (PICICT), pp. 13-17, 2021.

[24] C. Jagadeesh, Pravin R. Kshirsagar, G. Sarayu, G. Gouthami and B. Manasa, ”Artificial intelligence based Fake Job Recruitment Detection Using Machine Learning Approach”, Journal of Engineering Sciences, vol. 12, pp. 0377-9254, 2021.

[25] F. H. A. Shibly, Sharma Uzzal and H. M. M. Naleer, ”Performance comparison of two class boosted decision tree snd two class decision forest algorithms in predicting fake job postings”, 2021.

[26] Hridita Tabassum, Gitanjali Ghosh, Afra Atika and Amitabha Chakrabarty, ”Detecting Online Recruitment Fraud Using Machine Learning”, 2021 9th International Conference on Information and Communication Technology (ICoICT), pp. 472-477, 2021.

[27] J. Srinivas, K. Venkata Subba Reddy, G. J. Sunny Deol and P. Vara Prasada Rao, ”Automatic Fake News Detector in Social Media Using Machine Learning and Natural Language Processing Approaches” in Smart Computing Techniques and Applications, Singapore:Springer, pp. 295-305, 2021.

[28] G. Srinivas, A. Lakshmanarao, S. Sushma, M. V. Krishna and S. Neelima, ”Fake News Detection Using ML and DL Approaches,” 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT), Kollam, India, 2023, pp. 1322-1325, doi: 10.1109/ICCPCT58313.2023.10245398.