Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 16, Issue. 2: 2025
ISSN : **1906-9685**

AUTOMATED DETECTION OF DECEPTIVE JOB POSTINGS USING LSTM, BERT, AND CNN MODELS

SK.HIMAM BASHA[1], Bobbepalli.Sandhya[2]
#1. Assistant Professor, #2. Pursuing MCA,
Department of Master of Computer Applications
QIS COLLAGE OF ENGINEEING AND TECHNOLOGY
Vengamukkalapalem (V), Ongole, Prakasam dist, Andhra Pradesh- 523272

Abstract:

Most companies nowadays are using digital platforms for recruitment, but this has led to an increase in fraudulent job postings, resulting in significant financial losses for job seekers. To combat this issue, this paper proposes a deep learning-based approach for detecting online recruitment fraud (ORF) using a novel dataset comprising three source Fake Job Posting, Pakistan Job Posting, and US Job Posting. The proposed methodology utilizes Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pre-training Approach (RoBERTa) to convert job details into numerical vectors. Due to the high class imbalance in the dataset, the SMOTE (Synthetic Minority Over-sampling Technique) SMOBD variant is applied to balance the classes. Experimental results show that the combination of BERT features with SMOBD achieved the highest accuracy of 98.68% when integrated with a Convolutional Neural Network (CNN2D) for job classification. This approach effectively addresses the challenges posed by outdated datasets and enhances the detection of fraudulent job postings, contributing significantly to the fight against online recruitment.

Introduction:

In the age of advanced technology, the internet has drastically transformed our lives in different ways. The traditional way to do any activity has now been switched online. Therefore, seeking a job and hiring employees have also switched online. An online recruitment system (E- recruitment) is an internet application, the benefits of which encompass productivity, easiness, and efficacy. Most organizations prefer

online recruitment systems to provide job opportunities to potential candidates. Organizations publish job ads for their vacant positions through job portals, in which they mention job descriptions, including requirements, salary packages, offers, and facilities to be provided. Job seekers visit different online job advertising websites, seek job ads related to their interests, and apply for suitable jobs. The company then screens the CVs of applicants matching their requirements. The position is closed after fulfilling other formalities like interviewing and selecting potential candidates. The trend of posting online job advertisements was inflated during the global pandemic of COVID 2019. According to the World Economic Outlook Report, the International Monetary Fund (IMF) estimated that the unemployment rate increased to 13% at the peak time of the COVID-19 pandemic in 2020. These statistics were only 7.3% in 2019 and 3.9% in 2018. During the outbreak, many companies decided to post job openings online to provide facilities to job seekers. But, where a facility is provided to the public, it also allows online fraudsters to take advantage of their pessimism.

Literature Survey:

1) Detecting Fake Job Postings using Bidirectional LSTM — (arXiv / preprint)

Takeaway: Uses a Bi-LSTM on the Kaggle "Real / Fake Job Posting" dataset and reports strong sequence modeling benefits over baseline classical models. arXiv

Dataset: Kaggle "Real / Fake Job Posting" (~17–18K posts). Kaggle+1

Method: Text preprocessing → word embeddings (pretrained or trained), then a Bidirectional LSTM classifier (sometimes combined with metadata features). arXiv

Results: Bi-LSTM improves recall/F1 vs classical ML baselines in the paper (check specific numbers in the PDF). arXiv

Strengths / limits: Good at modeling sequential dependencies in job descriptions; limited by dataset imbalance and often simple handling of metadata (company profile, location). Opportunity: combine LSTM with attention or metadata fusion and more robust class-imbalance strategies. arXiv+1

2) Fraud-BERT: Transformer-based context-aware online recruitment fraud detection (peer/preprint)

Takeaway: Fine-tuning a BERT (or light variant) on job-ad text gives strong

contextual understanding and often outperforms shallow/deep RNNs on semantics and nuance. DNB+1

Dataset: Variants of the Kaggle dataset and other curated recruitment fraud corpora. DNB+1

Method: BERT (transfer-learning) fine-tuned for classification; some works propose domain-adapted BERT (Fraud-BERT) or DistilBERT for efficiency. DNB+1

Results: Transformer models show higher accuracy/F1 and better handling of subtle deceptive cues (phrasing, promises). DistilBERT/efficient variants trade a bit of accuracy for speed. DNB+1

Strengths / limits: Excellent semantic understanding; can be compute-heavy and may overfit small labelled sets. Opportunity: domain adaptation, model compression, and explainability (highlight which text spans drove the prediction). DNB

3) Hybrid CNN–LSTM / BERT–CNN / BERT–LSTM architectures applied to deception detection (adapted from fake-news / fraud literature)

Takeaway: Hybrid stacks (CNN for local n-gram feature extraction + LSTM for sequence; or BERT + CNN/LSTM) often outperform single models on deception/fraud classification because they capture both local patterns and longer context. (Examples shown in fake-news / fraud detection literature and adapted to job fraud.) PMC+1

Datasets / domains: Fake news and recruitment fraud corpora; several studies port these hybrids to job-ad detection, showing gains. PMC+1

Method: e.g., pretrained BERT → pooled token embeddings → CNN/LSTM head; or CNN over embeddings + LSTM on CNN outputs (stacked). PMC+1

Results: High accuracy in deception/fraud tasks; ensembling (BERT-LSTM + BERT-CNN) and late fusion can further boost results. TEM Journal+1

Strengths / limits: Combines strengths of both architectures; complexity and inference cost rise. Useful to explore lightweight stacks or ensemble distillation. PMC

4) Improving Fake Job Description Detection Using Deep Learning (NLP2FJD) — (peer-reviewed journal)

Takeaway: Presents an end-to-end deep-NLP pipeline for fake job description (FJD) detection and reports improved detection when combining textual deep features with metadata and engineered signals. Taylor & Francis Online

Dataset: Kaggle + extended/cleaned variants; authors emphasize improved preprocessing and metadata exploitation. Taylor & Francis Online+1

Method: Deep encoders (CNN / LSTM / transformer variants) + feature fusion with metadata (salary, location, company profile). Taylor & Francis Online

Results: Notable gains vs prior work, especially when metadata are properly encoded and used alongside text embeddings. Taylor & Francis Online

Strengths / limits: Good practical focus on data quality and feature fusion; may lack large-scale external validation. Opportunity: cross-platform datasets and temporal robustness testing. Taylor & Francis Online

5) Survey / Recent applied studies & small-scale projects (multiple conference / journal reports)

Takeaway: A number of recent applied papers and reports (IJPR, IJRPR, conference proceedings) demonstrate practical pipelines using DistilBERT, BERT, Bi-LSTM, CNN, and hybrid strategies on the same Kaggle benchmark — but they differ in preprocessing, imbalance handling, and evaluation rigor

Observation: Most work uses the same public Kaggle dataset (so be careful about overfitting to its quirks) and report metrics like accuracy/F1; fewer include robust cross-validation or external testbeds. Kaggle+1

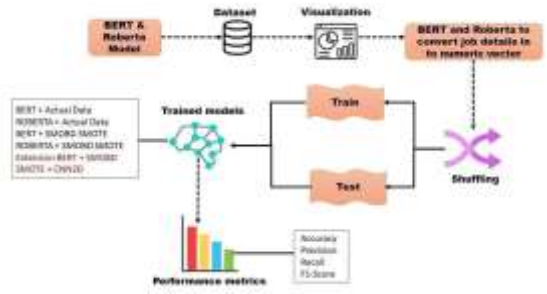
Gap: Lack of large, diverse, multi-platform labelled corpora and limited explainability analyses in many papers — opportunity for contributions (dataset curation, explainable classifiers, adversarial robustness).

In recent years, the growing use of online platforms for job recruitment has introduced both convenience and risks. While digital job portals offer accessible opportunities for job seekers and simplified hiring processes for employers, they have also become a target for fraudulent job postings. Online Recruitment Fraud (ORF) exploits the anonymity of the internet to mislead applicants, often resulting in financial loss and psychological stress.

This project focuses on building a deep learning-based system to effectively detect fraudulent job advertisements. Traditional systems that use classical machine learning algorithms—such as Naive Bayes, Decision Trees, and Logistic Regression—lack the sophistication to identify complex fraud patterns and often fail to handle class imbalance in datasets.

The proposed system addresses these challenges by leveraging advanced Natural Language Processing (NLP) models, specifically BERT (Bidirectional Encoder

System Architecture:

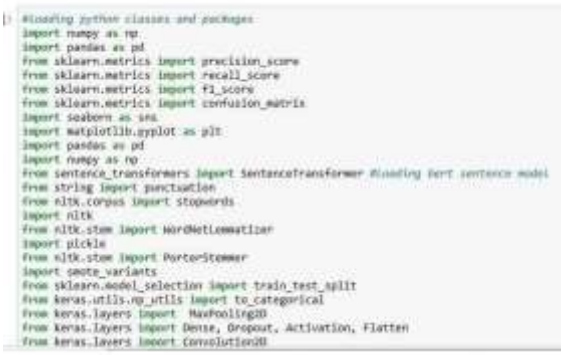


Representations from Transformers) and RoBERTa (Robustly Optimized BERT Approach). These models are capable of transforming job descriptions into meaningful numerical vectors that capture context and semantics more accurately than traditional methods. To resolve the issue of class imbalance—where real job postings significantly outnumber fraudulent ones—the SMOBD, a variant of the SMOTE (Synthetic Minority Over-sampling Technique), is applied. This ensures that minority class instances (fraudulent jobs) are adequately represented during training. Additionally, a 2D Convolutional Neural Network (CNN2D) is integrated to enhance feature extraction and boost the classification performance.

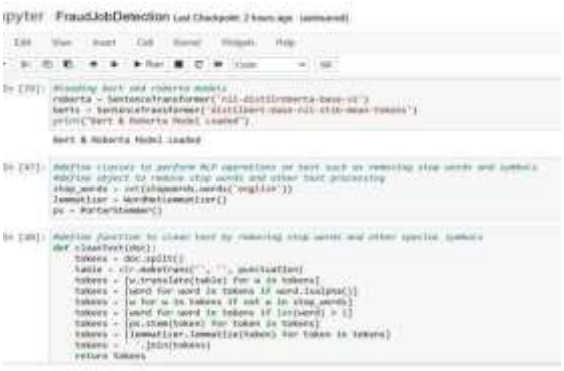
This system analysis outlines the shift from rule-based and shallow learning methods to deep learning architectures, supported by comprehensive datasets from multiple sources. It sets the groundwork for a robust, scalable, and secure solution capable of identifying fraudulent job postings in real-time, contributing to a safer and more trustworthy digital recruitment environment.

Implementation:

We have coded this project using JUPYTER notebook and below are the code and output screens with blue colour comments



In above screen importing required packages and classes

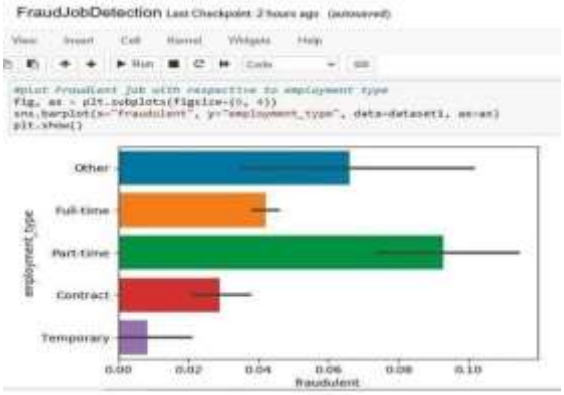


In above screen in first block creating and loading objects of BERT and Roberta models and then defining function to clean text data using NLP algorithms

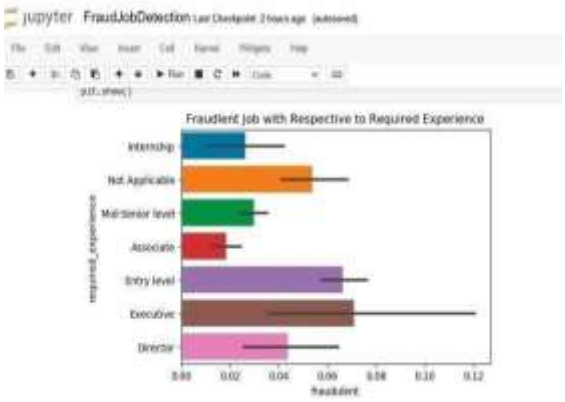
In above screen loading and displaying Fake job first dataset values

In above screen loading and displaying second dataset values.

In above screen loading and displaying 3rd dataset values

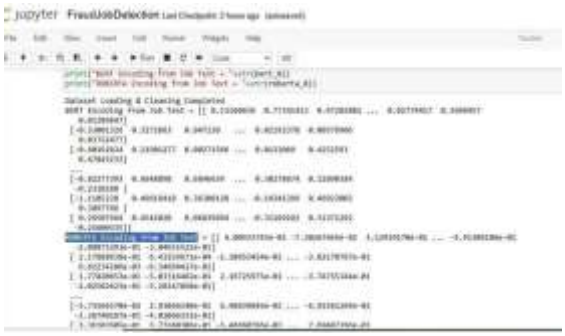


In above screen visualizing fraud job percentage in dataset with respect to employment type and in above graph can see 'Part Time' job type contains more fraud job. In above graph x- axis represents fraud job percentage and y-axis represents employment type

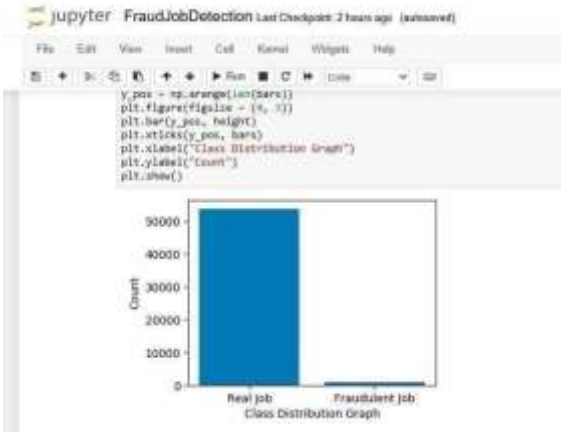


In above screen visualizing fraud job percentage with respect to 'Required Experience' and in above exploration we can say 'Executive' experience contains more fraud jobs

In above screen reading all job details from all 3 datasets and then input to BERT and Roberta to convert job details in to numeric vector and after executing above blocks will get below output

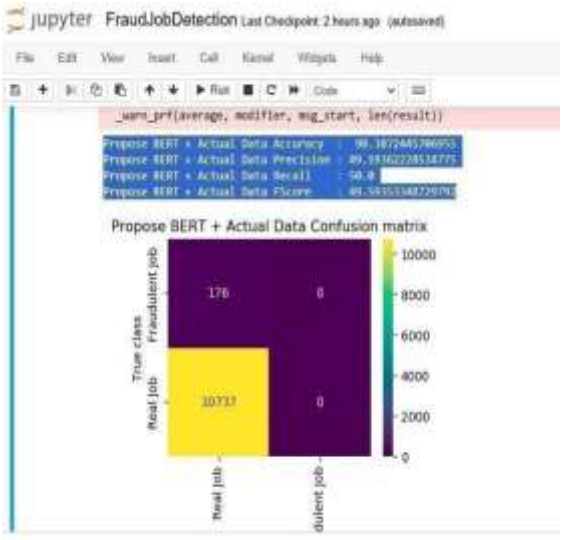


In above screen can see numeric vector generated from job text for both BERT and Roberta models

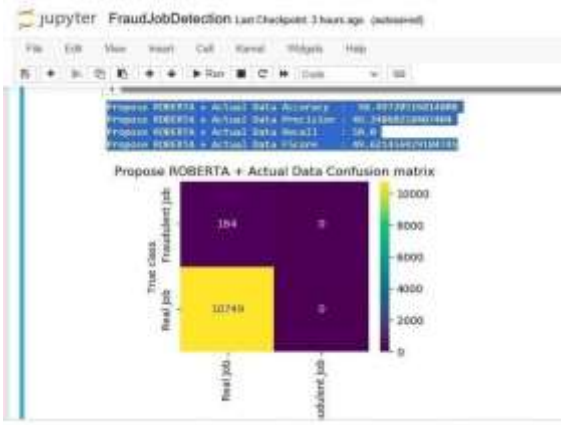


In above screen visualizing REAL and Fraud number of jobs exists in dataset and can see more jobs posting are REAL and very few records contains FAKE job posting and this having high imbalance data issue. In above graph x-axis represents Job Type and y-axis represents number of records

In above screen applying processing techniques such as shuffling and splitting data into train and test where 80% BERT and Roberta features will used for training and 20% for testing



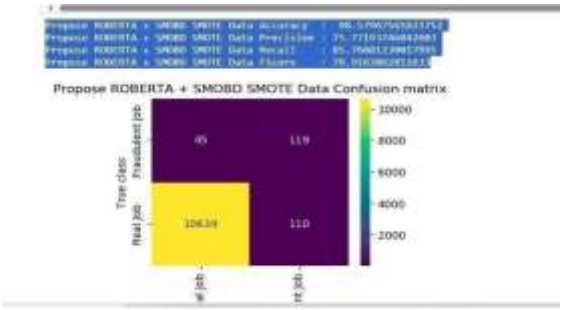
In above screen BERT neural network with actual data got 98% accuracy and can see other metrics like precision, recall and FSCORE. In confusion matrix graph x-axis represents Predicted Labels and y-axis represents true labels and then yellow and blue boxes in diagonal represents correct prediction count and remaining both blue boxes represents incorrect prediction count. In above graph can see without using SMOTE classifier predicted all jobs as REAL and fake count is 0 and this will happened because of imbalance data issue and can solve this issue using SMOTE invariants classes.



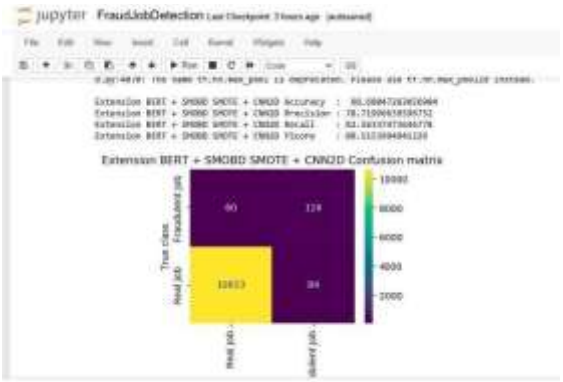
In above screen applying SMOTE SMOBD algorithm on both BERT and Roberta features and then in blue colour text can see

dataset contains 41000 records and in last line after applying SMOTE we can see dataset features size increase to 85000

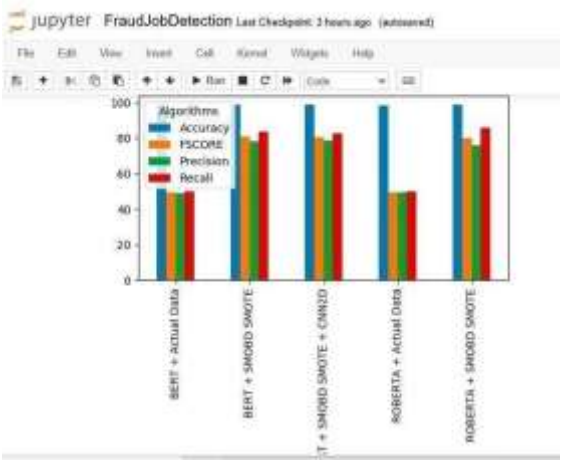
In above screen BERT + SMOTE got 98.66% accuracy and in confusion matrix graph we can see algorithm predicted both real and fraud jobs as no class has 0 predicted count which we see in previous confusion matrix without SMOTE



In above screen Roberta + SMOTE got 98.57% accuracy and can see this algorithm also predicted both Real and Fake jobs



In above screen extension got 98.68% accuracy which is higher than all propose algorithms and in confusion matrix can see both real and fake jobs predicted



In above screen displaying all algorithm performance comparison graph where x-axis represents algorithm names and y-axis represents accuracy and other metrics in different colour bars and in all algorithms Extension got high accuracy

Conclusion:

In conclusion, the proposed system for detecting online recruitment fraud (ORF) effectively addresses the increasing prevalence of fraudulent job postings on digital platforms. By integrating multiple advanced deep learning algorithms, including Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pre-training Approach (RoBERTa), the system enhances the capability to accurately identify fake job advertisements. The use of a novel dataset comprising postings from various sources, along with the application of the SMOTE SMOBD technique, significantly mitigates class imbalance issues, ensuring robust training and evaluation of the models. The results highlight that the combination of BERT features with SMOBD, when integrated with a Convolutional Neural Network (CNN2D), achieved the highest accuracy of 98.68%. This demonstrates the efficacy of the proposed system in distinguishing between genuine and fraudulent job postings. The project provides a valuable

framework that can help protect job seekers from online scams, ultimately contributing to a more secure recruitment process in the digital landscape.

Future Enhancements:

In future work, the project aims to further enhance the detection of online recruitment fraud by exploring additional machine learning techniques, such as ensemble methods and advanced feature extraction algorithms. Integrating recurrent neural networks (RNNs) and attention mechanisms may improve the model's ability to capture contextual information in job postings. Additionally, experimenting with transfer learning from pre-trained models can optimize performance on smaller datasets. These enhancements aim to refine accuracy and efficiency in identifying fraudulent job advertisements.

REFFERENCES

[1] G. Othman Alandjani, ‘‘Online fake job advertisement recognition and classification using machine learning,’’ 3C TIC, Cuadernos de Desarrollo Aplicados a las TIC, vol. 11, no. 1, pp. 251–267, Jun. 2022.

[2] A.Adhikari, A.Ram, R.Tang, and J.Lin, ‘‘DocBERT: BERT for document classification,’’ 2019, arXiv:1904.08398.

[3] I. M. Nasser, A. H. Alzaanin, and A. Y. Maghari, ‘‘Online recruitment fraud detection using ANN,’’ in Proc. Palestinian Int. Conf. Inf. Commun. Technol. (PICICT), Sep. 2021, pp. 13–17.

[4] C.Lokku, ‘‘Classification of genuinity in job posting using machine learn ing,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 12, pp. 1569–1575, Dec. 2021.

[5] S. U. Habiba, Md. K. Islam, and F. Tasnim, ‘‘A comparative study on fake job post prediction using different data mining techniques,’’ in Proc. 2nd Int. Conf. Robot., Electr. Signal Process. Techn. (ICREST), Dhaka, Bangladesh, Jan. 2021, pp. 543–546.

[6] Report Cyber. Accessed: Jun. 25, 2022. [Online]. Available: https://www.actionfraud.police.uk/

[7] S. Vidros, C. Kolias, G. Kambourakis, and L. Akoglu, ‘‘Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset,’’ Future Internet, vol. 9, no. 1, p. 6, Mar. 2017.

[8] S. Dutta and S. K. Bandyopadhyay, ‘‘Fake job recruitment detection using machine learning approach,’’ Int. J. Eng. Trends Technol., vol. 68, no. 4, pp. 48–53, Apr. 2020.

[9] B. Alghamdi and F. Alharby, ‘‘An

intelligent model for online recruitment fraud detection,'' J. Inf. Secur., vol. 10, no. 3, pp. 155–176, 2019.

[10] S. Lal, R. Jiaswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, ''ORFDetector: Ensemble learning based online recruitment fraud detec tion,'' in Proc. 12th Int. Conf. Contemp. Comput. (IC3), Noida, India, Aug. 2019, pp. 1–5.

[11] G. Kovács, ''An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets,'' Appl. Soft Comput., vol. 83, Oct. 2019, Art. no. 105662.

[12] S. Gazzah and N. E. B. Amara, ''New oversampling approaches based on polynomial fitting for imbalanced data sets,'' in Proc. 8th IAPR Int. Workshop Document Anal. Syst., Nara, Japan, Sep. 2008, pp. 677–684.

[13] O. Nindyati and I. G. Bagus Baskara Nugraha, ''Detecting scam in online job vacancy using behavioral features extraction,'' in Proc. Int. Conf. ICT Smart Soc. (ICISS), vol. 7, Bandung, Indonesia, Nov. 2019, pp. 1–4.

[14] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, ''Handling imbalanced datasets: A review,'' GESTS Int. Trans. Comput. Sci. Eng., vol. 30, no. 1, pp. 25–36, 2006.

[15] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, ''Toward credible evaluation of anomaly-based intrusion-detection methods,'' IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 40, no. 5, pp. 516–524, Sep. 2010.

[16] Y.-H. Liu and Y.-T. Chen, ''Total margin based adaptive fuzzy support vector machines for multiview face recognition,'' in Proc. IEEE Int. Conf. Syst., Man Cybern., Waikoloa, HI, USA, Oct. 2005, pp. 1704–1711.

[17] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, ''Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance,'' Neural Netw., vol. 21, nos. 2–3, pp. 427–436, Mar. 2008. 109406

[18] Y.Li,G.Sun,andY.Zhu,''Dataimbal anceproblemintextclassification,'' in Proc. 3rdInt. Symp.Inf. Process., Luxor, Egypt, Oct.2010,pp. 301–305.

[19] N.V.Chawla, K.W.Bowyer, L.O.Hall, and W.P.Kegelmeyer, ''SMOTE: Synthetic minority over-sampling technique,'' J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[20] S. Barua, M. M. Islam, and K. Murase, ‘‘ProWSyn: Proximity weighted synthetic oversampling technique for imbalanced data set learning,’’ in Proc. Pacific–Asia Conf. Knowl. Disc. Data Min. II, Gold Coast, QLD, Australia, Apr. 2013, pp. 317–328.

Authors:

Mr. Himambasha Shaik is an Assistant Professor in the Department of Master of Computer Applications at QIS College of Engineering and Technology, Ongole, Andhra Pradesh. He earned his Master of Computer Applications (MCA) from Anna University, Chennai. With a strong research background, He has authored and co-authored research papers published in reputed peer-reviewed journals. His research interests include Machine Learning, Artificial Intelligence, Cloud Computing, and Programming Languages. He is committed to advancing research and fostering innovation while mentoring students to excel in both academic and professional pursuits.

Ms. Bobbepalli Sandhya, currently pursuing Master of Computer Applications at QIS College of engineering and Technology (Autonomous), Ongole, Andhra Pradesh. She Completed Bsc(physics) from Sri Gowthami Degree College,Chirala, Andhra Pradesh. Her areas of interest are Machine learning & Cloud computing.