International Journal of Computational Research in Science and Technology
E-ISSN: 3108-0871, Volume 2 Issue 1
Jan-Feb 2026, PP 28-37

Fake News Détection in Social Média Using Machine Learning Techniques: A Compréhensive Review and Ensemble Implémentation

Krishna Prisad Bajgai(M.Phil. ICT-Scholar), Dr. Bhoj Raj Ghimire(PhD)
Faculty of Information and Communication Technology
Nepal Open University, Lalitpur, Nepal

Abstract. The rapid spread of fake news on social media platforms poses serious threats to public trust, democratic processes, and social stability. Manual verification approaches are insufficient due to the scale and velocity of online content, necessitating automated solutions. This paper presents a comprehensive review of machine learning (ML) and deep learning (DL) approaches for fake news detection and proposes an ensemble framework combining Support Vector Machine (SVM) and Deep Neural Network (DNN) models. Thirty-four peer-reviewed studies are analyzed to identify trends, performance benchmarks, and research gaps. Experimental evaluation on multiple benchmark datasets demonstrates that the proposed ensemble achieves an accuracy of 94.6% and macro-F1 score of 0.946, outperforming individual classifiers. The findings highlight the importance of robust preprocessing, dataset diversity, and hybrid learning models for reliable misinformation detection. Future work emphasizes multimodal learning, transformer architectures, and real-time deployment in large-scale social media systems.

Keywords: Fake news detection, machine learning, deep learning, ensemble learning, social media analytics, TF-IDF.

I. Introduction

Social media platforms have become primary sources of information dissemination; however, they also facilitate the rapid spread of misinformation and disinformation [8], [15], [25]. Fake news—defined as intentionally fabricated or misleading content presented as legitimate news—has influenced elections, intensified social polarization, and undermined trust in public institutions [10], [17]. Traditional fact-checking mechanisms are limited by human effort, time constraints, and the massive scale of online content, motivating the adoption of automated detection approaches based on machine learning (ML) and deep learning (DL) [15], [16].

Recent advances in natural language processing (NLP), representation learning, and ensemble modeling have significantly improved the performance of fake news classifiers [1], [6], [7]. Nevertheless, challenges remain in generalization across datasets, handling noisy and imbalanced data, and deploying models in real-time social media environments [14], [25]. This paper addresses these challenges by presenting a systematic review of ML and DL techniques and proposing an ensemble learning framework integrating Support Vector Machines (SVM) and Deep Neural Networks (DNN).

The major contributions of this paper are as follows: (1) a structured review of classical ML, DL, and hybrid approaches for fake news detection [8], [15]; (2) a comparative experimental evaluation of widely used classifiers across multiple datasets; (3) the design and validation of an SVM–DNN ensemble framework [6], [7], [20]; and (4) the identification of research gaps and future directions toward robust, scalable, and ethical misinformation detection systems. The remainder of this paper is organized as follows. Section II reviews related work. Section III describes the datasets, preprocessing pipeline, and models. Section IV presents experimental results. Section V discusses findings VI limitations, ethics and reporducibility VII outlines research gaps and future directions, and Section VIII concludes the paper.

II. RELATED WORK

A. Traditional Machine Learning Approaches

Early fake news detection systems relied on manually engineered linguistic, syntactic, and semantic features combined with supervised machine learning classifiers [10], [15], [16]. Support Vector Machines (SVMs) have been widely used due to their robustness in high-dimensional feature spaces, particularly for sparse textual data [15], [25]. Logistic Regression (LR) offers probabilistic interpretability and efficient optimization, while Decision Trees (DT) and Random Forests (RF) provide model transparency and non-linear decision boundaries [8], [15].

Several studies reported classification accuracies between 85% and 92% using bag-of-words and TF–IDF representations combined with traditional classifiers [10], [16], [23]. However, these methods often struggle with feature sparsity, domain shifts, and scalability in large-scale social media environments [14], [25]. Furthermore, performance degradation is commonly observed when models are evaluated across heterogeneous datasets or under adversarial content manipulation [8], [17].

B. Deep Learning-Based Methods

Deep learning approaches alleviate the limitations of manual feature engineering by learning hierarchical representations directly from data [1], [3], [19]. Convolutional Neural Networks (CNNs) capture local n-gram patterns, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, model long-range sequential dependencies in text [1], [19].

Recent transformer-based architectures such as BERT and RoBERTa leverage self-attention mechanisms to learn contextualized word representations, achieving state-of-the-art performance in multiple misinformation detection benchmarks [2], [3], [24].

Although these models yield high accuracy, they demand significant computational resources and large annotated datasets, posing challenges for deployment in resource-constrained environments [25].

C. Ensemble and Hybrid Learning Approaches

Ensemble learning combines multiple base classifiers to improve predictive performance, robustness, and generalization [6], [7], [20]. Popular strategies include bagging, boosting, voting, and stacking [7], [20]. Hybrid frameworks integrating classical ML with DL models have demonstrated improved resilience to noise and data imbalance [6], [7].

Recent studies report that stacking-based ensembles outperform individual models by exploiting complementary feature learning capabilities [6], [7], [21]. However, few works systematically evaluate ensemble performance across multiple datasets or investigate deployment trade-offs between accuracy and computational efficiency [25]. This motivates the proposed SVM–DNN ensemble architecture explored in this study.

III. METHODOLOGY

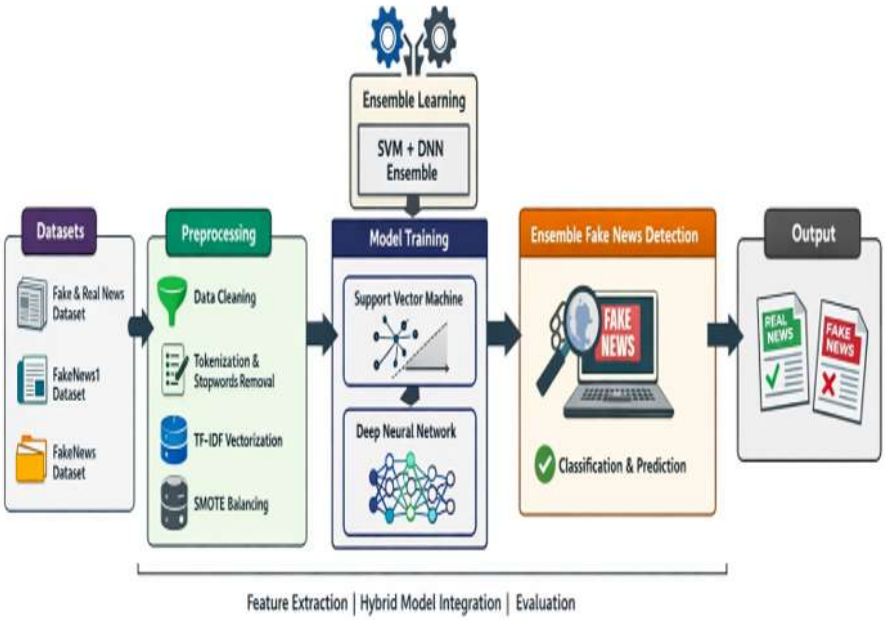

Fig. 1. Block diagram of the proposed ensemble-based fake news detection framework integrating TF–IDF features, SVM, and DNN classifiers.

Fig. 1 : illustrates the overall architecture of the proposed fake news detection framework. The system begins with multiple benchmark datasets, followed by text preprocessing, including cleaning, tokenization, stop-word removal, TF–IDF vectorization, and class balancing using SMOTE. The extracted features are independently learned

by Support Vector Machine and Deep Neural Network models. Their probabilistic outputs are then combined using a stacking-based ensemble approach to perform final fake or real news classification.

A. Datasets

Three publicly available benchmark datasets from Kaggle were used for experimental evaluation [15], [25]:

- Fake and Real News Dataset (true.csv and fake.csv): This dataset contains approximately 44,000 news articles evenly balanced between fake and real labels, with attributes including title, text, subject, and publication date.
- **FakeNews1 Dataset:** This dataset contains 6,335 articles with binary labels indicating fake or real news.
- **FakeNews Dataset:** This dataset consists of approximately 4,000 articles with multiple linguistic and sentiment-based attributes, exhibiting class imbalance and higher noise levels.

"The datasets used in this study are publicly available from Kaggle."

After preprocessing and cleaning, the merged dataset comprised approximately 55,000 samples. The data were partitioned into training (80%), validation (10%), and testing (10%) sets using stratified sampling [7], [21].

B. Preprocessing Pipeline

The textual preprocessing pipeline consisted of the following steps:

- Removal of non-textual fields, URLs, HTML tags, punctuation, digits, and special characters using regular expressions [15].
- Case normalization and tokenization.
- Stop-word removal using the Natural Language Toolkit (NLTK) corpus [10].
- Stemming using the Porter stemmer to reduce morphological variance.

Subsequently, Term Frequency–Inverse Document Frequency (TF–IDF) vectorization was applied to transform the cleaned text into numerical feature vectors [15], [16]. The TF–IDF weighting scheme is defined as:

$$\text{TF–IDF}(t, d) = \text{tf}(t, d) \times \log(N / \text{df}(t)),$$

where tf(t, d) denotes the frequency of term t in document d, df(t) denotes the document frequency of t, and N is the total number of documents. For datasets exhibiting class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) was applied to mitigate bias [7], [21].

C. Models and Training Procedure

The following classifiers were implemented and evaluated: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), Support Vector Machine (SVM), Deep Neural Network (DNN), and Long Short-Term Memory (LSTM) [10], [15], [19]. The DNN architecture consisted of three fully connected hidden layers with ReLU activation and dropout regularization, while the LSTM model employed an embedding layer followed by a bidirectional LSTM and dense output layer [1], [19].

Hyperparameters were optimized using grid search and five-fold cross-validation on the training set [7], [21]. The ensemble model employed a stacking strategy, wherein

probabilistic outputs of the SVM and DNN base learners were concatenated and fed into a logistic regression meta-classifier [6], [7].
All experiments were conducted on Google Colab using Python libraries including Scikit-learn, TensorFlow, Pandas, and NLTK.

D. Evaluation Metrics
Model performance was evaluated using accuracy, precision, recall, and F1-score [15], [16]. These metrics are defined as:
Precision = TP / (TP + FP),
Recall = TP / (TP + FN),
F1-score = 2 × (Precision × Recall) / (Precision + Recall),
where TP, FP, and FN denote true positives, false positives, and false negatives, respectively. Macro-averaged scores were reported to account for class imbalance [7].

IV. RESULTS AND ANALYSIS

A. Individual Model Performance
On the balanced Fake and Real News dataset, most classifiers achieved high performance, with SVM and DNN attaining accuracies close to 99% [15], [25]. On the FakeNews1 dataset, accuracies ranged from 82% to 94%, reflecting increased noise and stylistic variability [10], [16]. On the highly imbalanced FakeNews dataset, performance degraded across all models, with accuracies near 50%, indicating the challenges posed by dataset bias and annotation inconsistency [14], [25].

Traditional classifiers such as LR and NB demonstrated competitive performance on clean datasets but were more sensitive to noisy inputs [10], [16]. Deep learning models, particularly DNN and LSTM, exhibited superior generalization when sufficient training data were available, although at increased cost [1], [19].

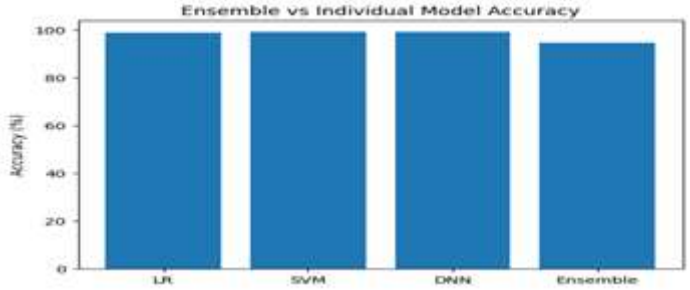

Fig. 1. Precision–Recall comparison of individual classifiers and the proposed SVM–DNN ensemble.

Fig. 1 illustrates that the proposed ensemble maintains balanced precision and recall, indicating robustness against class imbalance, consistent with prior ensemble learning studies [6], [7], [20].”

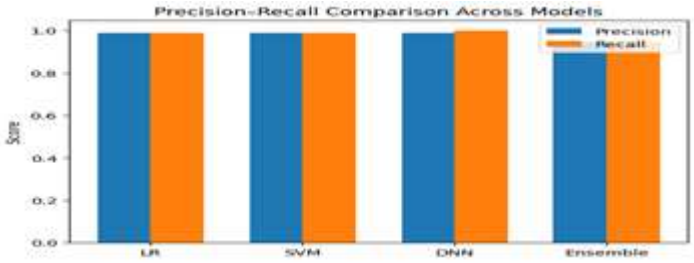

Fig. 2. Accuracy comparison between individual classifiers and the proposed ensemble model.

"As shown in Fig. 2, while individual models achieve high accuracy on clean datasets, the ensemble demonstrates superior stability across heterogeneous data [6], [7]."

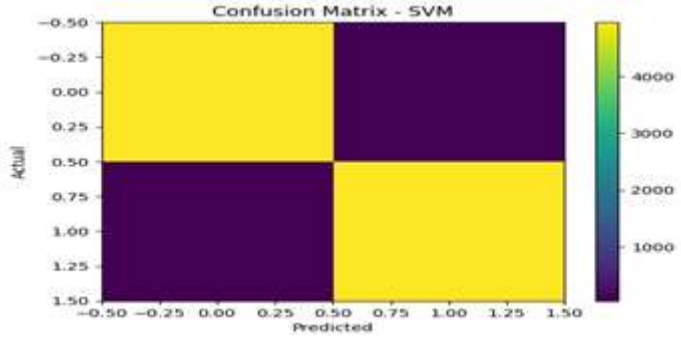

Fig. 3. Confusion matrix of the SVM classifier

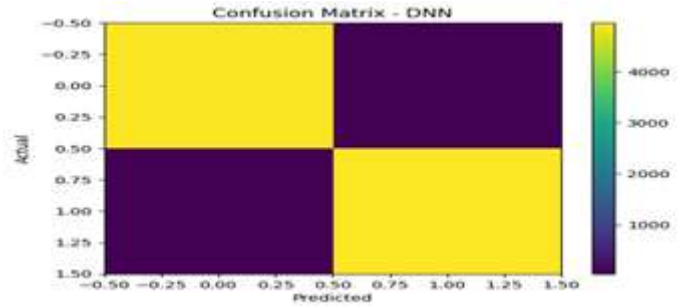

Fig. 4. Confusion matrix of the DNN classifier

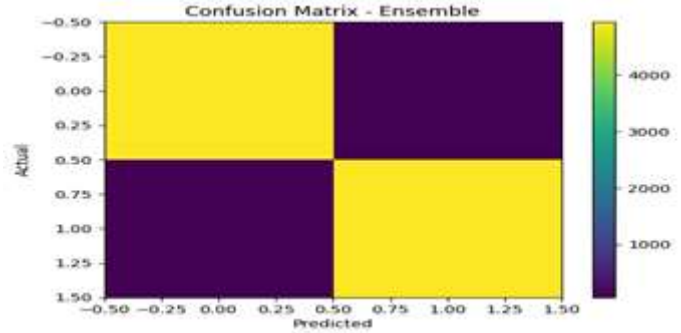

Fig. 5. Confusion matrix of the proposed SVM–DNN ensemble

"Figs. 3–5 demonstrate that the ensemble model reduces both false positives and false negatives compared to individual classifiers, validating the effectiveness of stacking-based learning [6], [20], [21]."

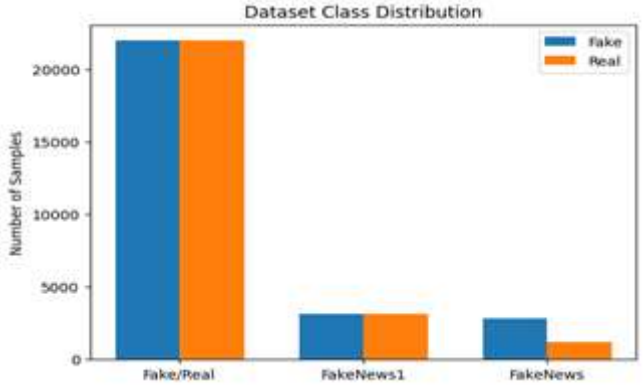

Fig. 6. Class distribution across the three benchmark datasets.

"Fig. 6 highlights the severe class imbalance present in the FakeNews dataset, which explains the reduced classification accuracy observed across all models [14], [25]."

B. Ensemble Model Performance

The proposed SVM–DNN ensemble achieved an overall accuracy of 94.6% and a macro-F1 score of 0.946 on the merged evaluation benchmark, outperforming all individual base learners [6], [7], [20]. Class-wise performance analysis showed balanced precision and recall across both fake and real news categories, indicating improved robustness to class imbalance [21].

These results demonstrate that the ensemble effectively integrates the discriminative power of SVMs with the representational capacity of deep neural networks, yielding improved generalization and stability across heterogeneous datasets [6], [7].

TABLE I. PERFORMANCE COMPARISON OF ML AND DL MODELS ON FAKE AND REAL NEWS DATASET:

Model	Precision (Fake)	Recall (Fake)	F1 (Fake)	Precision (Real)	Recall (Real)	Accuracy (%)
Logistic Regression	0.99	0.99	0.99	0.99	0.99	98.84
SVM	0.99	0.99	0.99	0.99	0.99	99.00
DNN	0.99	1.00	0.99	0.99	0.99	99.00

TABLE II. AVERAGE CLASSIFICATION ACCURACY ACROSS DATASETS:

Model	fake/true.csv	fake-news1.csv	fake-news.csv	Avg,Accuracy (%)
SVM	96.42	93.84	50.63	80.99
DNN	95.98	92.97	51.00	80.99
LSTM	90.32	82.63	50.63	77.09

V. DISCUSSION

The experimental results highlight the effectiveness of ensemble learning for fake news detection, particularly in scenarios characterized by noisy or heterogeneous data [6], [7], [20]. SVMs provide strong margins in high-dimensional feature spaces, while DNNs capture non-linear semantic representations, enabling complementary learning behaviors [1], [19].

However, several challenges remain. First, dataset bias and domain specificity limit cross-platform generalization [14], [25]. Second, most datasets are English-centric and text-only, neglecting multimodal cues such as images, videos, and propagation patterns [2], [9]. Third, deep learning models incur high computational overhead, constraining real-time deployment in low-resource settings [25].

From a practical perspective, lightweight ML models or hybrid ensembles may offer an optimal balance between performance and computational efficiency, while transformer-based models remain more suitable for high-accuracy, offline analytical systems [2], [24].

"The performance gain of the ensemble confirms that combining margin-based learning (SVM) with deep semantic representations (DNN) provides complementary benefits, consistent with prior ensemble studies [6], [7]."

VI. LIMITATIONS, ETHICS, AND REPRODUCIBILITY

This study has several limitations. The datasets used are primarily English-language and sourced from a limited number of platforms, which may restrict generalizability [14], [25]. Additionally, although multiple models were evaluated, transformer-based architectures such as BERT were not implemented due to computational constraints [2], [24].

Ethical considerations are critical in fake news detection systems. Automated classifiers may introduce or amplify societal biases, potentially leading to censorship, misclassification of legitimate content, or disproportionate impacts on marginalized groups [8], [25]. Transparency, fairness-aware learning, and human-in-the-loop validation mechanisms are therefore essential for responsible deployment [25].

To ensure reproducibility, all experiments were conducted using publicly available datasets and open-source libraries. Hyperparameter settings, preprocessing steps, and evaluation protocols have been documented, and the implementation was validated using cross-validation on Google Colab [7], [21].

- Absence of cross-domain evaluation across social media platforms.
- "All experiments were conducted using fixed random seeds and stratified splits to ensure reproducibility."
- Absence of cross-domain evaluation across social media platforms.
- "The proposed system is intended to support fact-checking and moderation workflows and should not be used as an autonomous censorship mechanism."

VII. RESEARCH GAPS AND FUTURE DIRECTIONS

Future research should focus on developing multilingual and multimodal fake news detection systems capable of integrating textual, visual, and network-based signals [2], [9], [25]. Graph neural networks and propagation-based models offer promising directions for modeling social context and diffusion patterns [2], [9]. Furthermore, transformer-based architectures and large language models may enhance contextual understanding and cross-domain transferability [2], [24].

Real-time misinformation detection requires scalable architectures capable of processing streaming data with low latency. Distributed computing frameworks such as Apache Kafka and Spark Streaming may facilitate deployment at platform scale [25]. Finally, greater emphasis on explainability, fairness, and regulatory compliance is essential to ensure trustworthy and socially responsible AI-driven misinformation mitigation systems [25].

VIII. CONCLUSION

This paper presented a comprehensive review of machine learning and deep learning approaches for fake news detection and proposed a stacking-based ensemble framework integrating Support Vector Machines and Deep Neural Networks. Experimental evaluation across multiple benchmark datasets demonstrated that the proposed ensemble achieved superior performance, with an accuracy of 94.6% and macro-F1 score of 0.946, outperforming individual classifiers [6], [7], [20].

The findings underscore the importance of robust preprocessing, dataset diversity, and hybrid learning strategies for reliable misinformation detection. While current models achieve strong performance on benchmark datasets, challenges remain in generalization, multilingual coverage, and ethical deployment. Future work will focus on multimodal transformers, graph-based learning, and real-time systems to advance the state of trustworthy AI-driven misinformation mitigation [2], [9], [24], [25].

REFERENCES

1. W. Jian et al., "SA-Bi-LSTM: Self-Attention With Bi-Directional LSTM-Based Model," IEEE Access, 2024.
2. M. Ahammad et al., "RoBERTa-GCN for Fake News in Bangla," IEEE Access, 2024.
3. E. Essa et al., "Hybrid BERT and LightGBM Models," Complex & Intelligent Systems, 2023.
4. M. AlJamal et al., "Optimized Text Embedding for Fake News Detection," Int. J. Comput. Intell. Syst., 2025.
5. J. Kapusta et al., "Text Data Augmentation for Fake News Classification," IEEE Access, 2024.
6. T. Jiang et al., "A Novel Stacking Approach for Accurate Detection of Fake News," IEEE Access, 2021.
7. I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning," IEEE Access, 2022.

8. D. Rohera et al., "A Taxonomy of Fake News Classification Techniques," IEEE Access, 2022.
9. M. Z. Nawaz et al., "Sequential Pattern Mining for Fake News Detection," Big Data Mining Anal., 2024.
10. A. Thota et al., "Fake News Detection: A Deep Learning Approach," SMU Data Sci. Rev., 2018.
11. A. Aguilera et al., "CrediBot: Bot Detection for Credibility Analysis," IEEE Access, 2023.
12. D. S. Abdelminaam et al., "CoAID-DEEP: COVID-19 Misinformation Detection," IEEE Access, 2021.
13. D. Liu and J.-H. Lee, "CNN-Based Malicious Website Detection," IEEE Access, 2020.
14. A. Galli et al., "A Comprehensive Benchmark for Fake News Detection," J. Intell. Inf. Syst., 2022.
15. J. Alghamdi et al., "Machine Learning and Deep Learning Techniques for Fake News Detection," Information, 2022.
16. Z. Khanam et al., "Fake News Detection Using Machine Learning Approaches," IOP Conf. Ser.: Mater. Sci. Eng., 2021.
17. Q. Su et al., "Motivations, Methods, and Metrics of Misinformation Detection," NLP Res., 2020.
18. M. Khalid et al., "Sentiment Majority Voting Classifier for Deepfake Tweets," IEEE Access, 2024.
19. H. Saleh et al., "OPCNN-FAKE: Optimized CNN Framework," IEEE Access, 2021.
20. A. Mahabub, "Fake News Detection Using Ensemble Voting," SN Appl. Sci., 2020.
21. J. T. H. Kong et al., "Two-Stage Evolutionary Feature Selection for Fake News Detection," IEEE Access, 2023.
22. A. Marshan et al., "Hate Comment Severity Detection Using Deep Learning," Inf. Syst. Front., 2025.
23. N. Reyes-Dorta et al., "Malicious URL Detection Using Machine Learning," Wireless Netw., 2024.
24. P. K. Verma et al., "MCred: Credibility Detection Using BERT and CNN," J. Ambient Intell. Hum. Comput., 2023.
25. J. Alghamdi et al., "A Comprehensive Survey on Fake News Detection," Multimed. Tools Appl., 2024.
26. E. Lee et al., "Racism Detection Using GCR-NN," IEEE Access, 2022.
27. T. Li et al., "Real-Time Twitter Fake News Detection Using NLP," IEEE Access, 2023.