DESIDOC Journal of Library & Information Technology, Vol. 45, No. 5, September 2025, pp. 439-447, DOI : 10.14429/djlit.21138

A Comparative Study of Machine Learning, Natural Language Processing and Hybrid Models for Academic Paper Acceptance Prediction: From Reviews to Decisions

Chandra Shekhar Pandey[$], Shriram Pandey[#,*], Tejash Pandey[@], Shweta Pandey[!], Harish Pandey[%] and Patanjali Mishra[^]

[$]*Department of Education, Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha - 442 001, India*
[#]*Department of Library and Information Science, Central University of Haryana, Mahendragarh - 123 031, India*
[@]*Guru Gobind Singh Indraprastha University, Delhi - 110 078, India*
[!]*GSV Central Library, Chhatrapati Shahu Ji Maharaj University, Kanpur - 208 024, India*
[%]*Department of Education, Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha - 442 001, India*
[^] *Department of Education, University of Allahabad, Allahabad – 211 002, India*
[*]*E-mail: shriram@cuh.ac.in*

ABSTRACT

The exponential increase in submissions to top-tier conferences and journals has placed unprecedented strain on editorial systems. To address this challenge, the present study explores the potential of computational modelling for predicting paper acceptance decisions based on peer review content as textual input as well as confidence score and recommendation score as numerical input in the models. We utilised the PeerConf dataset by Hasan, *et al.* which contains 3,242 reviews across 1,236 papers. In the study we design and evaluate three modelling approaches, including traditional ML models, transformer-based and sentiment-integrated NLP models (BERT, DistilBERT), and a novel hybrid model incorporating structured features, textual inputs and sentiment within ML pipelines. We have used accuracy and F1 scores to capture and compare the predictive effectiveness of the models. Python 3.10 environment and scikit-learn library were used for machine learning models, and Hugging Face Transformers v4.x was used for transformer-based models. The study contributes to the understanding of how hybrid models compare with ML and NLP-based models and provide a viable solution to predict the paper acceptance decisions. All models were trained in a GPU-enabled environment using PyTorch and Scikit-learn. The study also suggests the viability of different approaches for designing editorial support systems. We found that hybrid models outperformed ML and sentiment-integrated NLP models with 83.51 % accuracy and an F1 score of 72.91 %.

Keywords: Peer review automation; Academic paper acceptance prediction; Machine learning models; Transformer based models; Hybrid models

1. INTRODUCTION

With the overwhelming growth of scientific literature, the peer review process remains at its heart for ensuring scientific rigour. The increased pressure of timely and accurate decision-making in the review process requires significant expertise and scientific manpower for reducing bias and making the decision-making procedure unbiased and objective. Use of the latest advancements in machine learning and artificial intelligence has significant implications for decision-making in scientific research. While on the one hand these methods are being utilised for objectivity, bias reduction and efficiency in decision-making, there have always been concerns about their relative performance and accuracy. This study applies machine learning, NLP and Hybrid models for assessing the accuracy in prediction of peer review outcomes. Academic paper acceptance prediction on the basis of peer reviews and metadata has emerged as a critical research domain within AI-assisted scholarly communication Kang[1], *et al.*

Machine learning, NLP and Hybrid models are being applied for automating review procedures and predicting peer-review outcomes. A study by Xie[2] utilised linear regression for predicting institution-level acceptance based on historical data, previous acceptance rate and reviewer overlap. It is one of the earliest studies which laid the foundation for macro-level research analytics using ML for forecasting institutional research output. Skorikov[3], *et al.* applied machine learning with a random forest approach using structured features from paper metadata to predict acceptance, which demonstrated moderate accuracy. Bao[4], *et al.* used a machine learning-based frequent itemset mining approach for identifying common decision rules in reviews, using interpretable mining, which demonstrated clear interpretability and good accuracy. Bharti[5], *et al.* modelled review texts utilising attention mechanisms for improved predictions with the help of Deep Attention Networks with high accuracy.

Received : 14 March 2025, Revised : 05 July 2025
Accepted : 12 July 2025, Online published : 22 August 2025

Ghosal[6], *et al.* applied sentiment-augmented deep learning, combining sentiment analysis of reviews with deep learning models, leading to a 29 % error reduction compared to the baseline model. Pendyala[7], *et al.* compared machine learning algorithms, such as Support Vector Machine (SVM), with LLM models such as BERT, DistilBERT and PaLM. They found that the machine learning approach of SVM + TF-IDF was outperformed by LLM-based embeddings and neural networks.

Wang[8], *et al.* found that Large Language Models (LLMs) are highly vulnerable to adversarial attacks, as small edits in titles and abstracts can easily mislead the outcomes. Niu[9], *et al.* applied a machine learning model for predicting peer-review outcomes in the context of cybersecurity. They trained the Doc2Vec-based classifier machine learning approach with an accuracy of 91.2 %, using full paper text, and 83.02 % using abstracts. In contrast, ChatGPT scored only ~50-55 % accuracy, which is close to random guessing.

Machine learning, LLMs and hybrid models are being equipped with new analytical strategies and tools, and the pace of research in this area is unprecedented, with upcoming breakthroughs in artificial intelligence.

This study is situated at the intersection of these advancements and aims to conduct a comparative evaluation of ML, NLP, and hybrid models for academic paper acceptance prediction, using a curated peer review dataset that leverages TF-IDF-based text features and aspect scores from the PeerConf dataset Hasan[10], *et al.*

1.1 Research Question

How do traditional machine learning, transformer-based, and hybrid sentiment-integrated models compare in predicting academic paper acceptance decisions based on accuracy and F1-score, and which among them offers the best performance for potential use in automating peer review decisions?

1.2 Rationale of the Study

Editorial and peer review systems are experiencing increased demands and pressure with the rapid expansion of academic publishing and subsequent rise in manuscript submissions. Although peer review has been a gold standard of scholarly validation, it suffers from variability in quality, reviewer bias, and inconsistency in decision-making. The study builds upon the works of previous scholars who demonstrated that ML, NLP and hybrid approaches have potential for predicting final acceptance on the basis of numerical aspect data and textual reviews (Kang, *et al.* 2018; Kumar, *et al.* 2022; Bharti, *et al.* 2023; Hossain, *et al.* 2022)[11-14]. The study has a utilitarian rationale in the sense that it provides a logical and unbiased decision-making framework about peer-review outcomes.

The design uses a tripartite framework for predicting academic paper acceptance. The first approach uses a traditional ML pipeline incorporating TF-IDF vectorisation of peer review text alongside normalised recommendation and confidence scores. XGBoost, Random Forest, Linear SVC, and Naive Bayes classifiers are used to process the structured inputs of our dataset. This approach prioritises explainability and computational efficiency, reflecting real-world editorial settings where transparent decisions are necessary. The second approach uses transformer-based models (BERT, DistilBERT), fine-tuned with sentiment scores derived from VADER sentiment analysis for identifying reviewer tone and sentiment based on contextual language representations Kumar[15], *et al*. Thus, the second approach uses transformers and sentiment analysis for joint modelling capable of capturing textual semantics (the meaning and context) as well as the reviewer sentiment.

The third and novel dimension of this study is the development of a hybrid approach which acknowledges the importance of structured prediction by traditional ML models using recommendation and confidence scores as numeral inputs as well as TF-IDF and sentiment scores from textual data. This approach bridges the gap between structured feature modelling and affective signal capture, which is largely unexplored in prior literature.

2. LITERATURE REVIEW

2.1 Paper Acceptance Prediction Using ML/LLMs

Niu[16], *et al.* assessed machine learning models and ChatGPT for predicting top-tier computer conference peer review acceptance outcomes. They found the absence of a publicly available dataset of peer-review decisions and constructed their own dataset for negative samples by approximation using three heuristic rules of long preprint timelines, delayed and lower-tier publication. The Doc2Vec embedding technique was used to convert full papers or abstracts into high-dimensional vectors, and 14 ML models were evaluated, including Linear SVM (91.8 % accuracy, 0.918 F1 score), Logistic Regression (91.2 % accuracy, 0.912 F1 score), Gaussian Process (91.0 % accuracy, 0.910 F1 score), Voting Classifier Ensemble (91.0 % accuracy, 0.910 F1 score), and Linear Discriminant Analysis (90.3 % accuracy, 0.903 F1 score). The accuracy of ChatGPT-4 was 50-55 %, and the F1 score was 0.50, which demonstrated its inconsistency and randomness in predicting acceptance decisions. It was found that machine learning models can predict decisions for computer security papers with high accuracy, but ChatGPT lacks domain depth and fine-grained judgement and is vulnerable to hallucinations and biases.

Raja[17], *et al.* proposed a BERT-based classifier for automated recommendation score prediction on the basis of review text alone, as a precursor to final acceptance. It aimed to reduce reviewer subjectivity and address reviewer harshness for creating a standardised model-driven framework. Data preprocessing involved cleaning and tokenisation, standard preprocessing and creation of text-label pairs. The bert-base-uncased model was fine-tuned on the classification head (softmax over n

score classes). BERT can extract embeddings from the text of papers or reviewer comments and feed them into a prediction layer that helps in determining the recommendation scores.

Liu[18], *et al.* applied the LMC Bert model, which combines large language models with contrastive learning and BERT (Bidirectional Encoder Representations) for predicting acceptance of peer-reviewed papers. The proposed method used LLMs to extract the core information in a structured manner, and further, this summary is processed through BERT to extract semantic features. With the help of Momentum Contrastive Learning (MoCo), papers were compared to each other for quality differences. The tested LMCBert model outperformed Standard BERT (4.6 % improvement), SciBERT (3.9 % improvement) and prompt engineering-based LLMs.

The review of publications on paper acceptance prediction using ML/LLMs demonstrates that ML models can predict paper acceptance with high accuracy. Traditional ML models (e.g., SVM, logistic regression, LDA) achieved approximately 90 % accuracy. Doc2Vec-based pipelines Niu, *et al.* have shown that simple embeddings + classic ML can outperform large models like ChatGPT in this domain. LLMs alone are not enough, and combining LLMs and ML (LMCBert) gives the best results Liu[19], *et al*.

Although the absence of a real dataset of peer-reviewed decisions for predicting outcomes has been a gap for most of the studies. These studies simulate acceptance decisions on the basis of publicly available metadata such as conference publication records and arXiv preprints, focusing on preprint status, lower-tier venue or delay in publication as a proxy for rejection. Our study utilises actual peer-review decision metrics such as confidence scores, aspect scores, and real review comments.

2.2 Meta-Review Generation and Review Aggregation

Earlier studies have addresses isolated components of the peer review process. For instance, Anjum[20],*et al.* focused on topic modelling for reviewer matching, while Ghosal[21], *et al.* employed sentiment features in deep learning architectures to predict review outcomes. However, these efforts typically rely on a single model or overlook systematic benchmarking.

Pradhan[22], *et al.* proposed MRGen, a deep learning-based decision support system for the scholarly peer review process for modelling acceptance decision prediction and meta-review generation. They used the dataset available on PeerRead by Kang[23], *et al*. For final decision prediction Peer review texts, recommendation scores, and reviewer confidence scores were utilised as input. The model architecture components used were convolutional layer, LSTM, Bi-LSTM and attention mechanism. Final prediction layer composed of a dense, fully connected layer with all the extracted features from the deep layers and scores for predicting acceptance. It provided ~85.8 % accuracy for decision prediction.

Presented DeepMetaGen as an unsupervised deep learning model for generating template-based meta-reviews from peer review texts. It leveraged Aspect-Based Sentiment Analysis (ABSA) to extract key opinions across pre-defined aspects of the final decision template, such as clarity, originality, soundness, etc., and their associated sentiments. The system proposed four stages of work: aspect and sentiment classification (multitask deep neural network), opinion phrase extraction (rule-based approach using SpaCy with POS tagging), opinion filtering (TF-IDF scoring, cosine similarity, and SentiWordNet-based filtering), and summary generation (transformer-based generator model). The model was trained on the ReviewAdvisor dataset and outperformed ROUGE, BLEU, and BERTScore evaluations. As input to generate the meta-review, it utilised peer review texts, aspect categories, Sentiment Labels, Annotated Dataset, and gold meta-reviews (250 samples). The best-performing model was BERTScore, with the F1 score capturing about 55.34 % semantic similarity to human-written meta-reviews.

Hasan[24], *et al.* automated meta-review generation with a two-part system for decision and summarisation using ML + BERT for decisions and the T5 approach for summarisation on a constructed dataset of scrapped meta-reviews and a customised pipeline. This model automated both tasks using machine learning and NLP. Acceptance decision prediction used traditional ML classifiers (e.g., Decision Tree, SVM) using BERT embeddings and review scores. For meta-review generation, a transfer learning approach was used as a strategy of text summarisation with a fine-tuned T% transformer model. Used input features were paper title, abstract, peer reviews, recommendation scores and reviewer confidence scores. The paper title, abstract, and peer reviews were directly sourced from the PeerRead dataset and additional data collected from OpenReview, both of which provide openly accessible metadata and review content for scientific submissions. It was found that Random Forest and Decision Tree + BERT gave the best performance with 88.9 % accuracy and outperformed MRGen by ~3 % for acceptance decision prediction. For meta-review generation, the fine-tuned T5 transformer model outperformed other popular models like BERT, GPT-2, BART and Pegasus.

Zhao & Zhang[25] conducted a review of Survey Reviewer Assignment Algorithms (RAAs) developed from 1992 to 2022 in peer review systems. They have categorised these algorithms into three states: reviewer database construction, paper-reviewer matching computation and optimisation of reviewer assignment. It was found that the construction of the databases was performed generally by online recruiting, expert knowledge, or dynamic web crawling (e.g., TPMS). The matching computation was generally performed by subjective bidding, text-based similarities (TF-IDF, LDA, etc.) and multi-source information (e.g., social networks, authority). Optimisation of algorithms for reviewer assignment was performed by retrieval-based (per-paper) and matching-based (global assignment using ILP, max flow, heuristics) methods. NLP was indicated as a core enabler in the review by Zhao & Zhang.

Fiorillo & Mehta[26] examine how Artificial Intelligence (AI) can be integrated into the editorial workflow of scientific journals to accelerate manuscript review. ChatGPT was fine-tuned for peer-review tasks on a corpus of anonymised peer-review reports and editorial guidelines drawn from multiple journals to cover varied disciplines. Supervised learning on human-annotated reviews was used for fine-tuning, and Reinforced Learning from Human Feedback (RLHF) was utilised to refine response quality and tone. This finetuned model of ChatGPT leveraged OCR-based PDF ingestion to deliver line-level and thematic critiques in seconds.

Steingard[27], *et al.* explored the way to use generative AI to improve the evaluation of responsible business research for promoting the social impact of business. They proposed that AI can codify standards if trained on ethical frameworks like the SDGs and RRBM principles to evaluate papers consistently. They emphasised the importance of identifying responsible business research and building an AI tool called ChatSDG+RR7 on the GPT-4 platform, which was trained on UN Sustainable Development Goals (SDGs) and RRBM's seven principles. This model was evaluated by comparing AI evaluations vs. human editor decisions across 463 papers, and it modelled agreement for acceptance with 85% agreement and rejections with 100 % agreement.

Pendyala[28], *et al.* introduced a comprehensive system that integrates reviewer assignment, aspect score prediction, and paper acceptance classification. Their approach combines classical ML models such as Support Vector Machines (SVM) and Artificial Neural Networks (ANN) with large language models (LLMs) like BERT and PaLM 2, achieving up to 86 % test accuracy using DistilBERT embeddings. While their deployment-orientated architecture is novel, the study offers limited comparative analysis across traditional ML algorithms and does not deeply explore model interpretability or optimisation techniques.

Earlier studies focus on either traditional ML models or LLMs individually (e.g., Niu, *et al.* ; Liu[29-30], *et al.*) and don't combine structured prediction using traditional ML models and unstructured features like sentiment analysis as we have proposed in our hybrid model development and testing. Many of the studies used proxy data for modelling the acceptance Niu[31], *et al*. Although Pendyala[32], *et al.* used some hybrid features, they lacked model interpretability and hyperparameter optimisation, which limits their real-world application. There is limited comparison across LLMs, NLPs and Hybrid approaches in the literature, as we propose prediction using multiple features and approaches in our study. The novel dimension of this study is the development of a hybrid approach which acknowledges the importance of structured prediction by traditional ML models using recommendation and confidence scores as numeral inputs as well as TF-IDF and sentiment scores from textual data. This approach bridges the gap between structured feature modelling and affective signal capture, which is largely unexplored in prior literature.

3. METHODOLOGY

This study aims to compare traditional machine learning models, transformer-based NLP models and hybrid models (integrating sentiment analysis) for predicting paper acceptance decisions. We used the PeerConf dataset Hasan[33], *et al.* which contains 3,242 peer reviews (1,019 accepted and 2,223 rejected papers) for 1,236 academic papers drawn from multiple scientific conferences. The dataset structure is numerical and textual. Numerical fields include recommendation score and confidence score. The textual field contains unstructured review text and the final decision (accept/reject).

Final decisions were binarised into labels (1 = accept, 0 = reject). We used advanced cleaning (lowercasing, removing non-alphanumeric tokens and stopwords, and lemmatising words) to produce a "clean review" text field for feature extraction. We split the dataset into an 80 % training set and a 20 % test set. To prevent data leakage, we applied the feature engineering only on the training data and replicated it on the test data.

All training was conducted in a GPU-enabled environment using PyTorch and Scikit-learn. The final models were deployed on the test set for cross-approach evaluation.

3.1 Traditional ML Models (TF-IDF + Reviewer Scores)

For traditional machine learning models, textual data (review comments) was transformed using TF-IDF vectorisation, and numeric features included the recommendation score and confidence score, which were normalised to maintain scale consistency. These combined features were used as the input for a series of classification models, including Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB) and XGBoost Classifier. We performed hyperparameter tuning via grid search with cross-validation on the training set (e.g., regularisation strength C = 1.0 with L2 penalty for logistic regression; C = 0.5 and RBF kernel for SVM; number of estimators = 100 and max depth = 20 for random forest; number of neighbours = 5 and distance weighting for KNN; and maximum depth = 10 with entropy criterion for decision tree), after which the models were retrained on the full training set using the best-found parameters. We found that logistic regression (L2, C=1.0) achieved the highest cross-validation accuracy (~82.3 %), outperforming tuned SVM (C=0.5, kernel='linear'), random forest (n_estimators=100, max_depth=10), and naive Bayes (alpha=1.0) in hyperparameter tuning using grid search. Fig. 1 presents the ML pipeline structure.

3.2 Transformers Based NLP

State-of-the-art transformer language models were fine-tuned on the peer review texts for meeting objective 2. BERT (base, uncased) and DistilBERT (a lighter distilled version of BERT) were chosen as our base models. Each of these models were initialised with pre-trained weights

(Devlin, *et al.*, Liu, *et al.*)[34-35] and then fine-tuned for 3-5 epochs (batch size: 16, learning rate: 2e-5, optimiser: AdamW) on our training set for binary classification (accept vs. reject). HuggingFace Transformers were used for tokenisation, and we added a classification head on the [CLS] token. Fig. 2 presents the summary of the NLP pipeline.

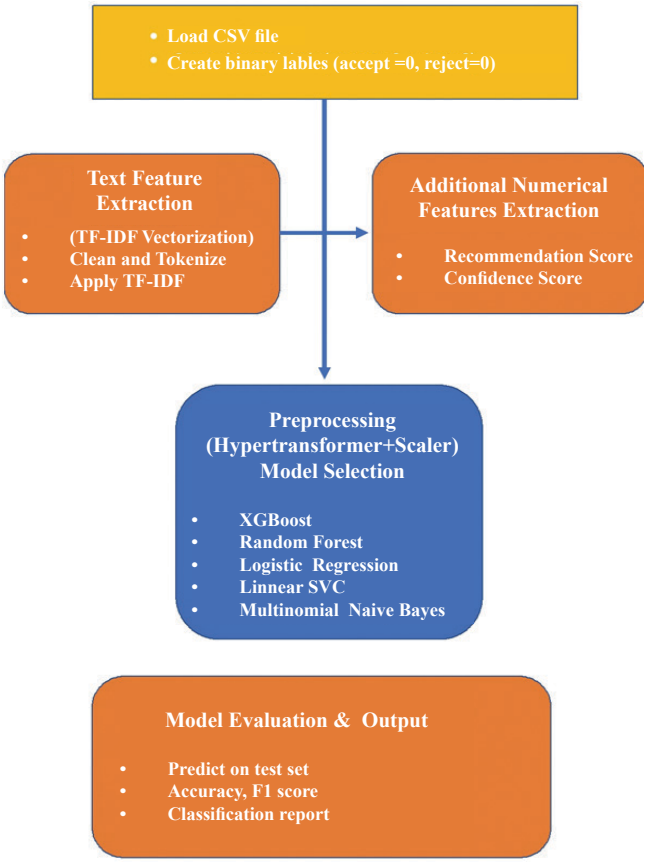

Figure 1. Traditional ML model pipeline.

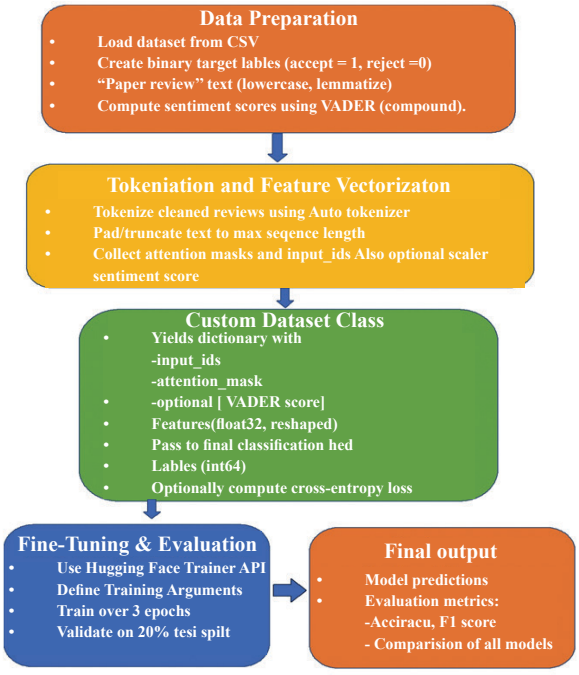

Figure 2. NLP pipeline.

The learning rate of models with the AdamW optimiser with a learning rate of 2e-5 for 4 epochs. Although we also used review text as input in a traditional ML model in objective 1, which was processed with TF-IDF, which works on a bag-of-words approach and doesn't understand the word, order, context or meaning. Although, in objective 2, the transformer-based approach uses tokenisation and embeddings (like from BERT) which capture context, semantics and word relationships.

3.3 Hybrid Sentiment-Integrated Models

This model is a fusion of traditional ML and transformer-based pipelines with emotional cues from sentiment analysis using the VADER sentiment analysis tool.

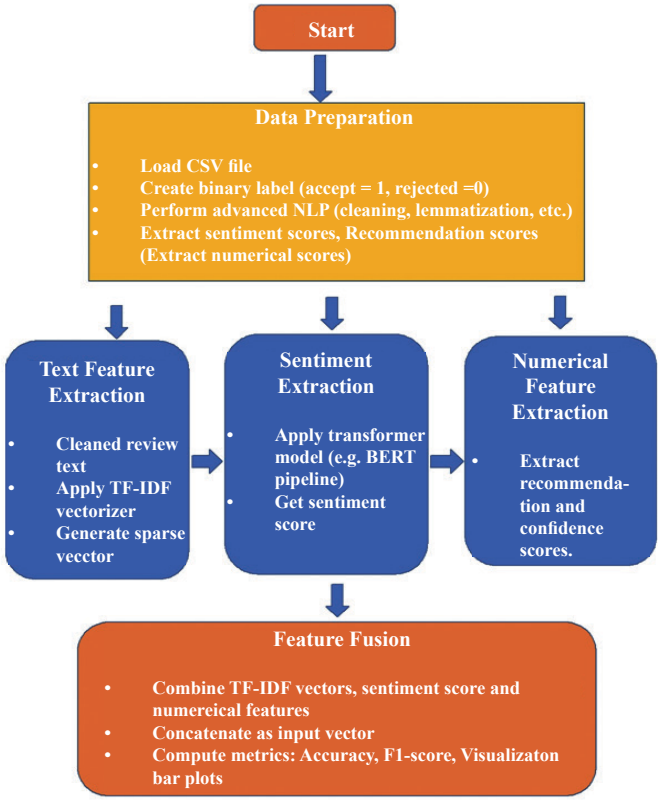

Figure 3. Hybrid model pipeline.

Traditional ML models (logistic regression and random forest) were retrained using a feature vector consisting of "TF-IDF features + recommendation score + confidence score + sentiment score". Figure 3 presents the Hybrid Model pipeline. To accommodate its decision boundary for the sentiment tone, we devised a custom architecture extending the BERT sequence classification model to accept an additional scalar sentiment feature concatenated to the model's pooled output. For enhancing standard transformer models like BERT and DistilBERT, we added an extra numerical sentiment score as input by modifying the final classification layer to accept this scalar sentiment feature (Devlin, *et al.*, 2019; Liu, *et al.* 2019; Sanh, *et al.* 2019). The final layers of these models expect a 768-dimensional vector, but we changed it to expect 769 inputs – 768 from the text (from the [CLS]

token) and 1 from the sentiment score. During training, each peer review text was tokenised and encoded using the Hugging Face Transformers pipeline Wolf, *et al*. The sentiment score was appended to the token's embedding, and the resulting 769-dimensional vector was passed into the new classification layer. These hybrid models were trained using the same hyperparameters as the base models with 3 epochs, a batch size of 16, a learning rate of 2e-5, and the AdamW optimiser (Loshchilov & Hutter, 2019).

3.4 Evaluation Protocol

To compare the predictive performance of all modelling approaches, we evaluated each model on a common hold-out test set using two standard metrics of accuracy and F1 score. The F1 score is useful in the context of imbalanced datasets Hasan[36], *et al.* such as those used in this study in which about 31 % of papers were accepted. To provide a fair comparison, we calculated F1-scores for the positive class (accepted papers) reflecting the model's ability to correctly identify accepted submissions, which is the focus of the study. All experiments were conducted in a Python 3.10 environment using the scikit-learn library Pedregosa[37], *et al.* for machine learning models and Hugging Face Transformers v4.x Wolf[38], *et al.* for transformer-based models.

4. RESULTS AND DISCUSSION

Table 1 and figure 4 present the comparative view of different models for their predictive performances. Traditional ML models such as Linear SVC and XGBoost achieved the highest test accuracy (82.13 %), which was followed closely by Logistic Regression (81.82 %). Linear SVC recorded the best F1 score (72 %) among the ML category, which suggests a good balance between precision and recall. However, MultinomialNB significantly underperformed in both accuracy (75.19 %) and F1-score (48 %), which indicates its limitations in handling the complexity of the peer review textual data. The transformer-based models (BERT and DistilBERT), which were fine-tuned with sentiment analysis, underperformed compared to traditional ML models. DistilBERT achieved slightly better accuracy (72.73 %) than BERT (72.11 %), but both exhibited low F1 scores in the range of 48.00 - 53.94 %.

Table 1. Comparative performance of models for paper acceptance prediction

Model	Test accuracy (%)	F1 score (%)	Notes
XGBoost	82.13	71.0	Traditional ML
Random forest	80.43	66.0	Traditional ML
Logistic regression	81.82	70.0	Traditional ML
Linear svc	82.13	72.0	Traditional ML
Multinomialnb	75.19	48.0	Traditional ML
Bert + sentiment	72.11	53.94	Transformer-based
Distilbert + sentiment	72.73	50.14	Transformer-based
Hybrid: logistic regression	83.51	72.91	Best performer
Hybrid: random forest	81.20	64.33	Hybrid

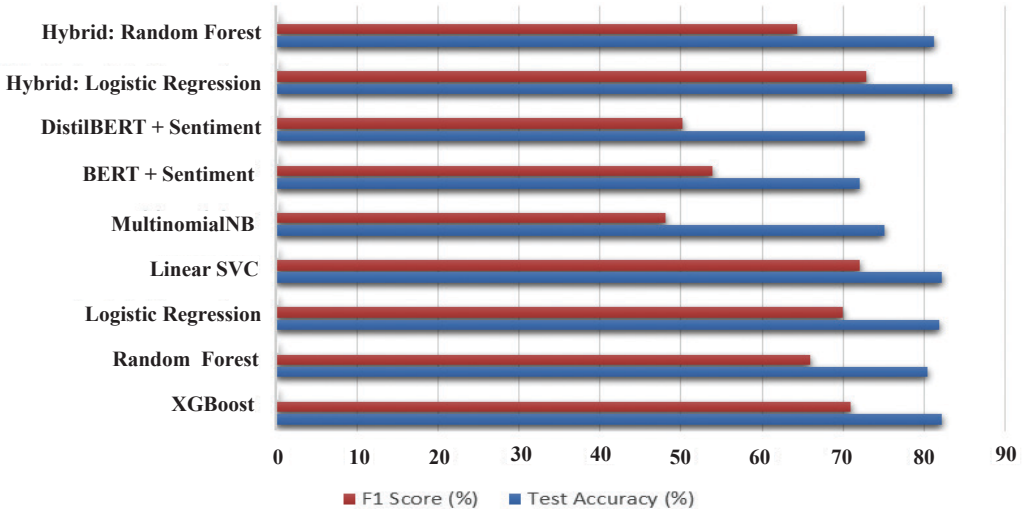

Figure 4. Comparative performance of models for paper acceptance prediction.

These findings are aligned with prior research by Niu[39], *et al.* who demonstrated that classical models such as linear SVM and logistic regression combined with effective textual embeddings such as Doc2Vec can outperform more complex models, including LLMs, in the context of peer review outcome prediction. Similalry Boukhris & Zaâbi[40] found that reviewer scores are highly predictive because they encapsulate the reviewer's judgement succinctly and often correlate directly with the final editorial decision. Hasan[41], *et al.* also found that SVMs trained on real peer review data in combination with review scores can match or exceed the performance of deeper models in decision support systems.

Hybrid models using logistic regression on the input of "TF-IDF features + recommendation score + confidence score + sentiment score" emerged as the best overall performer, achieving the highest accuracy (83.51 %) and F1 score (72.91 %). The second highest-performing model was the Hybrid Random Forest model, with an accuracy of 81.20 % but a slightly lower F1 score (64.33 %). Boukhris & Zaâbi[42] found that sentiment features enhance score-based models' predictability, while Riberio[43], *et al.* emphasise that sentiment captures soft cues and improves prediction when paired with scores.

5. CONCLUSION

We aimed to study the comparative effectiveness of traditional machine learning models, transformer-based sentiment-aware models and hybrid sentiment-integrated approaches in predicting academic paper acceptance decisions from peer review data from the PeerConf dataset. This dataset includes real-world reviewer recommendation scores, confidence scores, and review texts. Our results provided evidence that traditional machine learning models such as Linear SVC, XGBoost, and Logistic Regression perform remarkably well. Linear SVC and XGBoost demonstrated the best test accuracy (82.13 %), which was followed closely by Logistic Regression (81.82 %). These models achieved balanced F1 scores (~70–72 %). These findings align with the prior works by Niu[44], *et al.* & Boukhris & Zaâbi[45], which highlight the predictive value of structured reviewer inputs in peer review decision modelling. The hybrid logistic regression model integrating TF-IDF, recommendation score, confidence score, and sentiment into a unified feature set achieved the highest accuracy (83.51 %) and F1 score (72.91 %) among all tested approaches. The hybrid Random Forest model also improved over its non-hybrid version but was less effective in terms of F1 score. Sentiment integration in transformer models did not yield consistent improvements.

Collectively these findings emphasise the centrality of structured reviewer inputs, such as recommendation scores, in designing a practical system to support editorial decision-making. In our case, hybrid traditional models offered a low-complexity, interpretable and high-performance solution. Equipped with logistic regression's structured approach, TF-IDF-based review text processing and sentiment integration, this model outperforms other models. It proves to be a strong candidate for editorial support systems providing accuracy, interpretability, and computational efficiency.

Future work can expand to multi-conference or multilingual datasets, with aspect-based sentiment analysis, incorporation of additional models (DeBERTa, ELECTRA, or GPT-based architectures) and explainable AI (XAI) techniques such as SHAP or LIME to build trust in automated decisions by providing rationale for predictions. In traditional ML models, further bigram/trigram features can be added to TF-IDF, and SMOTE for imbalance can also be utilised. Multimodal inputs and integration of metadata (submission time, keywords, abstracts, and titles) and reviewer profiles (experience and expertise area) as features can be utilised for more accuracy and explainability.

REFERENCES

1. Kang JS, Ammar W, Schwartz R. A dataset and classifier for academic paper acceptance prediction. In: Proceedings of EMNLP. 2018.
2. Xie J. Predicting institution-level paper acceptance at conferences: A time-series regression approach. In: KDD Cup 2016 Proceedings. 2016.
3. Skorikov M, Momen S. Machine learning approach to predicting the acceptance of academic papers. In: Proceedings of the 2020 IEEE International conference on industry 4.0, artificial intelligence, and communications technology (IAICT); 2020:113–7. doi: 10.1109/IAICT50021.2020.9171996
4. Bao P, Hong W, Li X. Predicting paper acceptance via interpretable decision sets. In: Companion proceedings of the web conference 2021. New York: ACM; 2021:461-7. doi: 10.1145/3442442.3451363
5. PK, Ghosal T, Agarwal M, Ekbal A. PEERRec: An AI-based approach to automatically generate recommendations and predict decisions in peer review. Int J Digit Libr. 2023;25(1):55–72. doi: 10.1007/s00799-023-00333-1
6. Ghosal T, Verma R, Ekbal A, Bhattacharyya P. A sentiment augmented deep architecture to predict peer review outcomes. In: Proceedings of the 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL); 2019:414–5. doi: 10.1109/JCDL.2019.00067
7. Pendyala VS, Kamdar K, Mulchandani K. Automated research review support using machine learning, large language models, and natural language processing. Electronics. 2025;14(2):256. doi: 10.3390/electronics14020256
8. Wang Y, Li Z, Zhang X. Adversarial attacks on large language models using regularised gradients with continuous optimisation methods. Neurocomputing. 2023;550:11-20.
9. Niu L, Xue N, Pöpper C. Unveiling the sentinels: Assessing AI performance in cybersecurity peer review. arXiv. 2023. doi: 10.48550/arXiv.2309.05457

10. Hasan MT, Shamael MN, Billah M, Akter A, Hossain MAE, Islam S, *et al.* PeerConf: A dataset for peer review aggregation. Mendeley Data. 2022;V1. doi: 10.17632/wfsspy2gx8.1
11. Kang JS, Ammar W, Schwartz R. A dataset and classifier for academic paper acceptance prediction. In: Proceedings of EMNLP. 2018.
12. Kumar R, Jain A, Singh S. Integrating reviewer features and sentiment in BERT-based decision prediction. IEEE Trans Affect Comput. 2022;13(3):540–51.
13. Bharti R, Singh A, Garg N. Hybrid models for peer review classification: An empirical study. J Scholarly Publ. 2023;54(2):142–58.
14. Hossain M, Sengupta S, Shah N. Peer review analytics: Modeling tone, stance, and decision using deep learning. In: NeurIPS workshop on peer review. 2022.
15. Kumar S, Ghosal T, Ekbal A. DeepMetaGen: An unsupervised deep neural approach to generate template-based meta-reviews leveraging on aspect category and sentiment analysis from peer reviews. Int J Digit Libr. 2023;24:263–81. doi: 10.1007/s00799-023-00348-3
16. Niu L, Xue N, Pöpper C. Unveiling the sentinels: Assessing AI performance in cybersecurity peer review. arXiv. 2023. doi: 10.48550/arXiv.2309.05457
17. Raja K, Shanmugavadivel S, Sivaraja MSM. Harnessing BERT for the automation of peer review process by prediction of recommendation score. In: Edwards BI, Tanko BL, Klufallah M, Abuhassna H, Chinedu CC, editors. Reimagining transformative educational spaces. Singapore: Springer. 2024:365-74. doi: 10.1007/978-981-97-8752-4_20
18. Liu C, Zhang X, Zhao H, Liu Z, Xi X, Yu L. LMCBert: An automatic academic paper rating model based on large language models and contrastive learning. IEEE Trans Cybern. 2025. doi: 10.1109/TCYB.2025.3550203
19. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv. 2019. https://arxiv.org/abs/1907.11692
20. Anjum O, Gong H, Bhat S, Xiong J, Hwu WM. PARE: A paper-reviewer matching approach using a common topic space. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th International joint conference on natural language processing (EMNLP-IJCNLP). 2019:518–28. doi: 10.18653/v1/D19-1050
21. Ghosal T, Verma R, Ekbal A, Bhattacharyya P. A sentiment augmented deep architecture to predict peer review outcomes. In: Proceedings of the 2019 ACM/IEEE Joint conference on digital libraries (JCDL). 2019:414–5. doi: 10.1109/JCDL.2019.00086
22. Pradhan T, Bhatia C, Kumar P, Pal S. A deep neural architecture based meta-review generation and final decision prediction of a scholarly article. Neurocomputing. 2021;428:218–38. doi: 10.1016/j.neucom.2020.11.004
23. Kang JS, Ammar W, Schwartz R. A dataset and classifier for academic paper acceptance prediction. In: Proceedings of EMNLP. 2018.
24. Hasan S, Ghosal D, Das D. Sentiment-aware BERT for academic decision modeling. In: Proceedings of ACL 2024.
25. Zhao X, Zhang Y. Reviewer assignment algorithms for peer review automation. Inf Process Manage. 2022;59(5):103028
26. Fiorillo L, Mehta V. Accelerating editorial processes in scientific journals: Leveraging AI for rapid manuscript review. Oral Oncol Rep. 2024;10:100511. doi: 10.1016/j.oor.2024.100511
27. Steingard DS, Reibstein D, Normandin M. Harnessing Generative AI to drive responsible business research and accelerate social impact. The wharton school research paper. 2024 Nov 11. https://ssrn.com/abstract=5016963 doi: 10.2139/ssrn.5016963
28. Pendyala VS, Kamdar K, Mulchandani K. Automated research review support using machine learning, large language models, and natural language processing. Electronics. 2025;14(2):256. doi: 10.3390/electronics14020256
29. Niu L, Xue N, Pöpper C. Unveiling the sentinels: Assessing AI performance in cybersecurity peer review. arXiv. 2023. doi: 10.48550/arXiv.2309.05457
30. Liu C, Zhang X, Zhao H, Liu Z, Xi X, Yu L. LMCBert: An automatic academic paper rating model based on large language models and contrastive learning. IEEE Trans Cybern. 2025. doi: 10.1109/TCYB.2025.3550203
31. Niu L, Xue N, Pöpper C. Unveiling the sentinels: Assessing AI performance in cybersecurity peer review. arXiv. 2023. doi: 10.48550/arXiv.2309.05457
32. Pendyala VS, Kamdar K, Mulchandani K. Automated research review support using machine learning, large language models, and natural language processing. Electronics. 2025;14(2):256. doi: 10.3390/electronics14020256
33. Hasan MT, Shamael MN, Billah M, Akter A, Hossain MAE, Islam S, *et al.* PeerConf: A dataset for peer review aggregation. Mendeley Data. 2022;V1. doi: 10.17632/wfsspy2gx8.1
34. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT; 2019:4171-86. https://aclanthology.org/N19-1423/
35. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, *et al.* RoBERTa: A robustly optimised BERT pretraining approach. arXiv. 2019. https://arxiv.org/abs/1907.11692
36. Hasan MT, Shamael MN, Billah M, Akter A, Hossain MAE, Islam S, *et al.* PeerConf: A dataset for peer

review aggregation. Mendeley Data. 2022;V1. doi: 10.17632/wfsspy2gx8.1
37. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825-2830.
38. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, *et al.* Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020:38–45.
39. Niu L, Xue N, Pöpper C. Unveiling the sentinels: Assessing AI performance in cybersecurity peer review. arXiv. 2023. doi: 10.48550/arXiv.2309.05457
40. Boukhris I, Zaâbi C. A GAN-BERT based decision making approach in peer review. Soc Netw Anal Min. 2024;14(1):107.
41. Hasan R, Farabi SF, Al Mahmud MA, *et al.* Applying the technology acceptance model (TAM) in information technology system to evaluate the adoption of decision support system. J Comput Commun. 2024;12(8):a323-a330
42. Boukhris I, Zaâbi C. A GAN-BERT based decision making approach in peer review. Soc Netw Anal Min. 2024;14(1):107
43. Ribeiro JMC, Alburquerque E, Viana B. Acceptance decision prediction in peer-review through sentiment analysis. In: Proceedings of the 2021 Brazilian Conference on Intelligent Systems (BRACIS); 2021 Nov 29-Dec 2; São Paulo, Brazil. Piscataway (NJ): IEEE; 2021. p. 289-294.
44. Niu L, Xue N, Pöpper C. Unveiling the sentinels: Assessing AI performance in cybersecurity peer review. arXiv. 2023. doi: 10.48550/arXiv.2309.05457
45. Boukhris I, Zaâbi C. A GAN-BERT based decision making approach in peer review. Soc Netw Anal Min. 2024;14(1):107

CONTRIBUTIONS

Dr. Chandra Shekhar Pandey is Assistant Professor at Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha. His research interests include: Educational psychology, Philosophy, and Psychometrics.
He contributed to the methodology, validation, original draft writing, and review and editing of the manuscript.

Prof. Shriram Pandey is Professor and Head, Department of Library and Information Science, Central University of Haryana, Mahendragarh. His research areas include: Educational technology, E-learning, Semantic web, and Knowledge management.
He contributed to methodology, review and editing, investigation, formal analysis, visualisation, and supervision.

Mr. Tejash Pandey is a BTech. student in Artificial Intelligence and Machine Learning at Guru Gobind Singh Indraprastha University, Delhi. His research focuses on Computer vision, Natural language processing, and Recommendation systems.
He contributed to methodology, review and editing, investigation, and data cleaning.

Dr. Shweta Pandey is Deputy Librarian at Chhatrapati Shahu Ji Maharaj University, Kanpur. Her expertise lies in Library services, Digitisation, and Academic management.
She contributed to review and editing, and formal analysis.

Dr. Harish Pandey is an Assistant Professor at Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha. His research interests include: Mathematics education, Ethnomathematics, and AI in education.
He contributed to review and editing, and visualisation.

Dr. Patanjali Mishra is Associate Professor at the Department of Education, University of Allahabad. His research covers Language teaching, Research methodology, and Educational foundations.
He contributed to the original draft writing and review and editing of the manuscript.