I.J. Modern Education and Computer Science, **2025, 6, 146-159**
Published Online on December 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2025.06.10

Leveraging Convolutional Neural Network to Enhance the Performance of Ensemble Learning in Scientific Article Classification

I. Nyoman Switrayana*
Faculty of Engineering, Bumigora University, Mataram, Indonesia
Email: nyoman.switrayana@universitasbumigora.ac.id
ORCID iD: https://orcid.org/0009-0006-0771-5537
*Corresponding Author

Neny Sulistianingsih
Department of computer science, Master program, Bumigora University, Mataram, Indonesia
Email: neny.sulistianingsih@universitasbumigora.ac.id
ORCID iD: https://orcid.org/0000-0003-0548-5038

Received: 11 November 2024; Revised: 26 February, 2025; Accepted: 02 May, 2025; Published: 08 December, 2025

Abstract: The classification of scientific articles faces challenges due to the complexity and diversity of academic content. In response to this issue, a new approach is proposed, utilizing Ensemble Learning, specifically Decision Tree, Random Forest, AdaBoost, and XGBoost, along with Convolutional Neural Network (CNN) techniques. This study utilizes the arXiv dataset, comparing the effectiveness of Term Frequency-Inverse Document Frequency (TFIDF) and Sentence-BERT (SBERT) for text representation. To further refine feature extraction, vectors derived from SBERT are integrated into the CNN framework for dimensionality reduction and obtaining more representative feature maps named latent feature vectors. The study also observes the impact of incorporating both the title and abstract on performance, demonstrating that richer textual information enhances model accuracy. The hybrid model (CNN + Ensemble Learning) demonstrates a substantial improvement in classification accuracy compared to traditional Ensemble Learning. The evaluation shows that CNN + SBERT with XGBoost achieved the highest accuracy of 94.62%, showcasing the benefits of combining advanced feature extraction techniques with powerful models. This research emphasizes the potential of integrating CNN within the Ensemble Learning paradigm to enhance the performance of scientific article classification and provides insights into the crucial role of CNN in improving model accuracy. Additionally, the study highlights the superior performance of SBERT in feature extraction, contributing beneficially to the overall model.

Index Terms: Scientific Article, Classification, Convolutional Neural Network, Ensemble Learning, TFIDF, SBERT

1. Introduction

The classification of scientific articles has become a critical challenge in the realm of Natural Language Processing (NLP), particularly due to the exponential increase in published literature each year. Automated classification plays a vital role in aiding researchers to discover relevant articles, organize information, and enhance research efficiency. However, manually classifying millions of articles is an arduous and time-consuming task. Text classification is a method used to categorize various types of documents into target categories based on selected features and document content [1,2]. Consequently, the advent of machine learning techniques has emerged as a significant focus of research aimed at addressing this challenge. Various methods, such as neural networks, including Convolutional Neural Networks (CNN), have demonstrated considerable efficacy in improving text classification performance, especially in multi-label classification problems [3]**.** For instance, [4] proposed a hierarchical discourse approach utilizing CNN to tackle data imbalance through Top-K based resampling, while [5] introduced label distribution prediction techniques to enhance the initial classification outcomes generated by baseline classifiers. Moreover, [6] provided a comparative analysis of several text classification algorithms, highlighting the superior accuracy of Support Vector Machines (SVM) in classifying scientific abstracts.

Despite substantial progress in the development of automated classification methods, challenges such as data imbalance and low classification accuracy in certain categories persist. Ensemble learning approaches, which combine multiple classification models, have been employed to mitigate these issues; however, the integration of CNN within ensemble frameworks requires further exploration to augment scientific article classification performance. The capability of CNNs to extract essential features from textual data at both word and sentence levels, combined with ensemble learning techniques, is posited to yield significant advancements in classification outcomes. Previous studies underscore the importance of supervised machine learning methods for large-scale text classification. For example, [7] demonstrated the effectiveness of supervised learning in classifying AI-related patents, highlighting its superiority over traditional keyword-based approaches. Additionally, [8] examined the application of CNN and other neural network techniques in text classification, revealing that the integration of neural networks with ensemble learning results in more efficient classification performance. Other notable works, such as those by [9] on ensemble learning for disease prediction and [10] on discrete wavelet transforms in EEG signal classification, further illustrate the evolving landscape of machine learning applications in classification tasks. Furthermore, recent studies stress the necessity of leveraging both title and abstract information to enhance model performance. [11] highlighted that the inclusion of both title and abstract allows for a more nuanced understanding of the content, which is essential for effective categorization, especially in fields like AI and medical research. Previous work by concatenating the title and abstract has shown significant potential in improving performance in paper recommendation systems [12]**.** By utilizing pre-trained transformer models, their approach demonstrated the importance of comprehensive input data in maximizing classification accuracy. This finding is highly relevant to the current study's goal of examining how richer textual input can improve model performance when compared to relying solely on the title. Incorporating both elements enriches the context, leading to more accurate document classification and better generalizability across diverse datasets.

In light of these developments, this research seeks to leverage CNNs to strengthen the performance of ensemble learning methods such as Decision Trees, Adaboost, Random Forest, and XGBoost in classifying scientific articles. Specifically, the study aims to address three primary research questions:

Q1: How does the performance improve when using input text consisting of both the title and abstract compared to using only the title? This question explores the impact of incorporating richer textual information on model performance.

Q2: What is the comparative effectiveness of Term Frequency-Inverse Document Frequency (TFIDF), an independent contextual feature extraction method, versus SBERT, which provides dependent contextual embeddings? This question evaluates the benefits of contextually aware features over traditional methods.

Q3: What role does CNN play in improving ensemble models through the dimensionality reduction of SBERT features? This question examines how CNN can enhance model performance by optimizing feature representation and dimensionality reduction.

In conjunction with these questions, [13] emphasizes the transformative role of AI in scholarly publishing, where ensemble models combined with neural networks have shown promising results in improving the readability and semantic richness of generated titles and abstracts. By using these methods, researchers have demonstrated that AI can outperform human authors in generating titles with higher relevance and accuracy, further affirming the importance of utilizing sophisticated machine learning approaches in the analysis and classification of textual data [13]. By investigating these questions, this study aspires to contribute to the advancement of more effective and accurate classification methods in the scientific domain, thereby addressing existing gaps in the literature and enhancing the overall utility of automated classification systems. The integration of CNNs with ensemble learning techniques, alongside the use of both title and abstract for text input, holds the potential to significantly improve the classification accuracy of scientific articles, benefiting researchers and expanding the applicability of these models in various fields.

The structure of this article is organized as follows: First section provides the introduction, discussing the background, gap analysis, and research objectives. Then, literature review section provides a detailed overview of related work in the field, highlighting key studies and methodologies employed in scientific article classification. Next section outlines the methodology used in this research, including the data collection process, preprocessing steps, and the architecture of the CNN-enhanced ensemble models. Result section presents the experimental setup and results, offering a comprehensive analysis of the performance metrics of the proposed models compared to traditional ensemble methods. Followed by next section which is discusses the findings, their implications, and potential limitations of the study. Finally, the article with a summary of the main contributions and suggestions for future research directions will be stated last.

2. Literature Review

Recent advancements in text classification have witnessed a surge in the application of deep learning models, each demonstrating varying degrees of success across different datasets and methodologies. The application of the gradient boosting model in study [14] demonstrated superior performance compared to other supervised learning models in the task of article classification. Research by [15] introduced BERT-BiGRU, achieving a notable precision of 89.5%,

surpassing conventional BERT models in accuracy. Meanwhile, [16] explored various hybrid deep learning architectures like HBLCNN with Glove and FastText embeddings, reporting accuracies ranging from 62.51% to 99.94%, with HBLCNN demonstrating superior performance. Similarly, [17] highlighted the effectiveness of Text-CNN and other CNN-based models in achieving superior results compared to traditional methods. Research by [4] applied Top-K resampling using BERT, achieving high precision, recall, and F1 scores with accuracies ranging from 0.951 to 0.993, emphasizing the method's robustness in complex data scenarios.

Moreover, [18] employed R-GAT on heterogeneous graph representations, demonstrating improved accuracy over MLP models, particularly achieving 70.11% accuracy in precision. Research by [19] introduced AttentionMFF, showcasing its superior accuracy of 92.42% and robust performance in identifying categories with precision and recall above 88%. [6] compared SVM, Naïve Bayes, K-Nearest Neighbor, and Decision Tree models using TF-IDF and BOW, highlighting SVM + TFIDF's dominance in precision and recall metrics. Furthermore, studies by [20] on web news datasets and [21] on Uzbek news datasets illustrated the efficacy of SVM and ensemble methods like Random Forest in achieving high accuracies of 95.04% and 86.88%, respectively. These findings underscore the versatility and applicability of machine learning models across diverse datasets and highlight ongoing advancements in optimizing model performance through novel architectures and data preprocessing techniques.

Despite the significant advancements highlighted in the literature, a distinct gap exists in the integration of Convolutional Neural Networks (CNNs) with ensemble learning methods for the classification of scientific articles. Previous studies, such as those exploring BERT-BiGRU and hybrid architectures like HBLCNN, have focused on standalone deep learning models, achieving notable accuracies but often lacking in feature extraction optimization when faced with the complexities of scientific texts. Moreover, while some studies have highlighted the performance of traditional machine learning models, such as SVM and ensemble techniques on specific datasets, there has been limited exploration of how CNNs can enhance the feature extraction process within ensemble frameworks. The study introduces a novel approach by integrating Convolutional Neural Networks (CNNs) with ensemble learning methods like Decision Tree, Adaboost, Random Forest, and XGBoost. This approach differs from conventional methods by combining the strengths of CNNs in feature extraction from textual data with the robustness of ensemble learning in improving classification accuracy and generalization. By training models on both title and abstract data using Sentence-BERT (SBERT) embeddings, the study aims to enhance classification accuracy by capturing semantic similarities and nuances embedded within scientific articles. This hybrid approach not only aims to surpass traditional methodologies in accuracy metrics but also seeks to optimize model performance through innovative architectural integrations and advanced data preprocessing techniques, thereby contributing to the evolving landscape of text classification research. The-state-of-the-art of this study can be seen on Table 1.

Table 1. The State-of-the-art

Paper	Extraction features model	Classification model	Evaluation model			
			Precision	Recall	F1Score	Accuracy
[15]	BERT	Bi-GRU	√			
[16]	Word embedding models	BLCNN, BLCNN2, RandBLCNNd, HBLCNN, HBLCNN2D				√
[19]	BERT	AttentionMFF, TextCNN, BERT	√	√	√	√
[4]	BERT	Top-K Resampling-Based Approach	√	√	√	
This research	- TF-IDF - SBERT	Hybrid model between CNN and ensemble learning	√	√	√	√

3. Methodology

To accomplish the research objectives, the methodology encompasses multiple stages, comprising data collection, preprocessing, feature extraction utilizing TF-IDF and SBERT, and modeling with two distinct scenarios aimed at improving classification accuracy. The final phase involves evaluating the model's performance. These research stages are depicted in Figure 1. The primary goal of this study is to boost ensemble learning performance through the integration of a hybrid CNN and ensemble learning model. This integration is expected to make a substantial contribution to enhancing the automated classification performance of scientific articles.

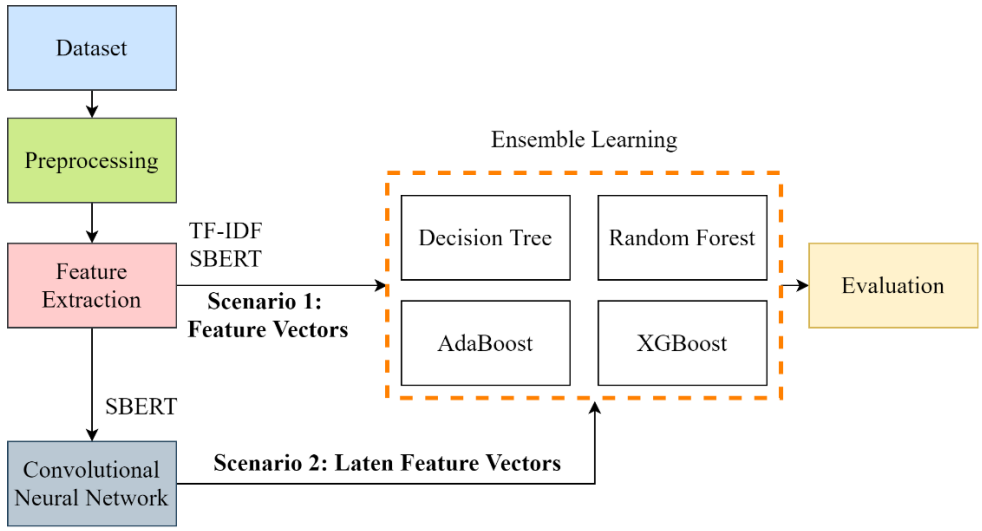

Fig. 1. Methodology overview for Scientific Article Classification

3.1 Dataset

The dataset used in this study is sourced from arXiv, a well-known repository of scientific papers that is publicly accessible and continuously updated. The specific version of the dataset used for this research is Version 153, dated November 13, 2023, formatted in JSON files. The dataset contains a total of 134,969 articles, covering a wide range of scientific disciplines, with 157 unique article categories. This dataset includes extensive metadata associated with each scientific paper, such as the ArXiv ID, submitter information, author details, paper title, additional comments like page count and figure details, journal reference, Digital Object Identifier (DOI), abstract, categories or tags within the ArXiv system, and a comprehensive version history. For this study, the features extracted and analyzed from the dataset include the paper title, abstract, and categories. These features were selected for their crucial role in enabling subsequent stages of data preprocessing, feature extraction, and modeling for scientific article classification. They provide fundamental information necessary for understanding and categorizing the content and context of each paper.

3.2 Preprocessing

The preprocessing steps involve tokenization, removing punctuation, removing stopwords, and converting text to lowercase. Tokenization is necessary to break the text into smaller, more manageable units (tokens), such as words. This step is particularly important as it prepares the text for feature extraction techniques like TF-IDF. Removing punctuation eliminates non-alphanumeric characters. Stopwords (e.g., "the", "is", "and", "to") are common words that occur frequently but carry little meaning in the context of classification tasks. Removing stopwords helps reduce the number of features in the dataset, thereby focusing the model on the more informative words. This process enhances computational efficiency and improves model accuracy by eliminating unnecessary noise. Converting text to lowercase ensures uniformity in word representation. While additional preprocessing steps such as stemming, lemmatization, or word normalization are common, they were not applied in this study. This decision is based on the nature of the dataset, which consists of scientific articles written in a formal and standardized style. These texts contain well-defined technical terms, and applying such steps could alter the meaning of specialized vocabulary. For instance, stemming or lemmatization might modify key terms, while word normalization could interfere with established scientific terminology. The preprocessing steps conducted can be observed in Figure 2. These steps collectively prepare text data for effective natural language processing and machine learning tasks, enhancing data quality and analysis outcomes.

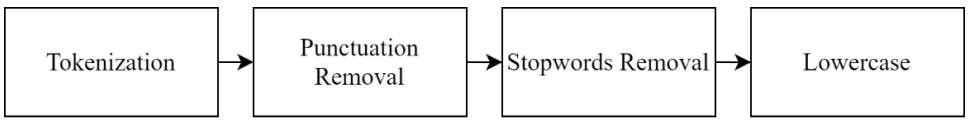

Fig. 2. Preprocessing phase

3.3 Feature Extraction

In the realm of natural language processing, the selection and application of feature extraction methods play a crucial role in enhancing the accuracy and effectiveness of text classification tasks. This study focuses on comparing the performance of context-independent and context-dependent feature extraction techniques: TF-IDF (Term Frequency-Inverse Document Frequency) and SBERT (Sentence-BERT), leveraging the state-of-the-art model, all-mpnet-base v2. TF-IDF, a classic statistical approach, evaluates the significance of words within a document corpus by combining term

frequency and inverse document frequency metrics [22]**.** It measures how often a term appears in a document while adjusting for its frequency across the entire corpus, thereby highlighting distinctive words crucial to each document's content. In contrast, SBERT utilizes Transformer-based architectures like BERT to generate dense, context-aware sentence embeddings [23]. BERT, while effective for word-level embeddings, was not ideal for this study due to its limitation in generating fixed-length sentence representations, which are essential for text classification tasks. SBERT overcomes this by fine-tuning BERT for sentence-level tasks, offering efficient, fixed-length embeddings [24]. Moreover, SBERT with the all-mpnet-base v2 model, currently one of the best performing in SBERT's documentation, provides superior sentence embeddings. This model strikes an optimal balance between performance and computational efficiency, making it ideal for large datasets and downstream classification tasks.

These embeddings capture subtle semantic relationships between words in sentences, offering a richer representation of textual data compared to traditional bag-of-words or TF-IDF approaches. In evaluating these methods, the study examines their effectiveness in enhancing ensemble learning models for scientific article classification. Due to SBERT's dense feature representation, these embeddings are further processed using a Convolutional Neural Network (CNN) to obtain latent feature vectors. These latent feature vectors are then utilized in ensemble learning to improve classification accuracy. A latent feature vector is a representation of the original data that has undergone reduction processes, minimizing its dimensions or complexity while retaining essential information [12]**.** This representation is valued for its ability to preserve the essence and meaning of complex data in a simplified format that facilitates easier processing and further analysis. In the context of using SBERT for scientific article classification, the latent feature vector generated after reduction using CNN

3.4 Modelling

Modeling in this study involves two primary scenarios to evaluate the performance of various feature extraction techniques in classifying scientific articles, considering the variation in input data. The input data variation includes using only the title feature and comparing it with using both title and abstract features in each scenario. The details of the research scenarios are presented in Table 2. In the first scenario, the approach begins by training an ensemble model using two main feature extraction methods: TF-IDF and SBERT. TF-IDF is utilized to provide numerical representations that depict the importance of key words in each document, while SBERT generates sentence embeddings that are rich in semantic meaning and context. In this scenario, the primary focus is to observe how these embedding representations contribute to the performance of the ensemble model in classifying scientific articles. SBERT embeddings, which capture more complex semantic relationships, are expected to enhance the accuracy and prediction precision of the ensemble model.

Table 2. Experimental scenario

Scenario 1				**Scenario 2**	
Training with title (TFIDF)	Training with title + abstract (TFIDF)	Training with title (SBERT)	Training with title + abstract (SBERT)	Training with title (SBERT)	Training with title + abstract (SBERT)
Decision Tree	Decision Tree	Decision Tree	Decision Tree	CNN + Decision Tree	CNN + Decision Tree
Adaboost	Adaboost	Adaboost	Adaboost	CNN + Adaboost	CNN + Adaboost
Random Forest	Random Forest	Random Forest	Random Forest	CNN + Random Forest	CNN + Random Forest
XGBoost	XGBoost	XGBoost	XGBoost	CNN + XGBoost	CNN + XGBoost

In the second scenario, a different approach is applied by integrating Convolutional Neural Network (CNN) after the feature extraction process using SBERT. CNN is employed to reduce the dimensionality of SBERT embeddings, producing more focused and reduced latent feature vectors. These latent feature vectors are then used as input in the ensemble model, aiming to improve representation and enhance classification capability. In this scenario, due to the vector dimension size of 768 generated by SBERT, which is then fed into CNN, it is reshaped into a 28x28 matrix with zero-padding for the remaining dimensions. The target dimensionality changes to 512 after passing through the convolutional layers and dense connections. This step allows the system to better capture complex patterns and deeper semantic relationships within scientific article texts. The proposed CNN architecture functions as an encoder, compressing the original SBERT embeddings into lower-dimensional latent representations. To assess how much information is preserved during this compression process, a decoder that mirrors the encoder architecture is implemented. This decoder reconstructs the original embeddings from the latent vectors. The reconstruction quality is evaluated using Mean Squared Error (MSE), which quantifies the average squared difference between the original and reconstructed embeddings. The MSE is computed using (1) [25]:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \quad (1)$$

Where x_i is the original value, $\hat{x}_i$ is the reconstructed value, and n is the number of features. This encoder-decoder architecture is also known as an autoencoder, a type of neural network designed to learn efficient representations of data by compressing (encoding) and then reconstructing (decoding) it while minimizing information loss. Both scenarios provide a comprehensive view of how the application of different feature extraction techniques can influence the performance of classification models. To ensure a fair and consistent evaluation, an 80:20 train-test split was applied, where 80% of the data was used for training the models and the remaining 20% was reserved for testing. By comparing the results from both scenarios, this study aims to provide valuable insights into the development of more advanced and effective automated classification systems for the scientific research domain.

3.5 Evaluation

Model evaluation is essential for assessing the effectiveness of classification models. Beyond overall accuracy, evaluating a classification model involves examining various metrics to understand how well it distinguishes between different categories. The confusion matrix is pivotal in this process, offering critical insights into model performance. Key values such as True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) are generated by classification models. TP indicates correctly predicted positive instances, while FP represents instances incorrectly predicted as positive when they are negative. FN counts instances incorrectly predicted as negative when they are positive, and TN counts accurately predicted negative instances. Common performance metrics include accuracy, precision, recall, and F1-score each providing specific measures of the model's classification capabilities. The calculation of each evaluation metric is calculated using (2-5) [2]:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{F1} - \text{Score} = \frac{2\ x\ Precision\ x\ Recall}{Precision+Recall} \tag{4}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

The use of accuracy, precision, recall, and F1-score in this study is intentional due to their complementary strengths. Accuracy provides an overall measure of performance but can be misleading in imbalanced datasets. Precision and recall help evaluate the model's ability to correctly identify relevant classes, particularly when false positives or false negatives carry different weights. The F1-score balances these two, making it especially valuable in multi-class and imbalanced classification tasks such as this study. These metrics are widely recommended in the evaluation of machine learning models for text classification and imbalanced data scenarios [26].

4. Results and Discussion

4.1 Dataset

The study concentrates on four specific categories of scientific articles: 'Quantum Physics', 'Materials Science', 'Statistical Mechanics', and 'Machine Learning'. These categories were chosen based on their top frequency and significance within the dataset, while also ensuring data variation across scientific fields. The composition of the dataset is illustrated in Figure 3. In this dataset, each article can have multiple tags or categories associated with it. However, for this study, only the primary category assigned to each article is utilized (the first tag). This study focuses on a curated subset of 29,084 articles, drawn from four major scientific categories: Quantum Physics (15,296 articles), Materials Science (6,318 articles), Statistical Mechanics (3,803 articles), and Machine Learning (3,667 articles). However, it is important to note that the dataset is inherently unbalanced, with a dominant number of articles in *Quantum Physics.*

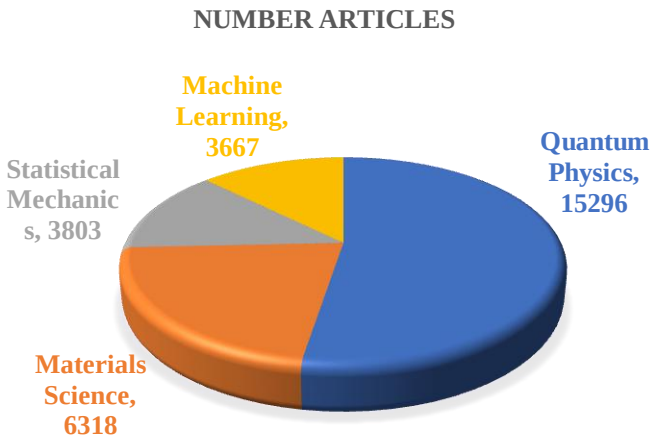

Fig. 3. Composition of categories in dataset

4.2 Preprocessing

The text preprocessing steps applied to the titles and the abtracts included tokenization, punctuation removal, stopwords removal, and conversion to lowercase. These processes are crucial for preparing text data for analysis, as they help standardize the input, reduce noise, and ensure that the models focus on the most relevant information. Each step's impact on the text data is detailed in Table 3, demonstrating the progressive refinement of the input titles as they undergo these preprocessing stages. This approach aims to enhance the quality of the text data, thereby improving the performance of subsequent machine learning models.

Table 3. Results of each process in the preprocessing steps

Input text	Tokenization	Punctuation removal	Stopwords removal	Lowercase
The Pursuit of Uniqueness: Extending Valiant-Vazirani Theorem to the Probabilistic and Quantum Settings	['The', 'Pursuit', 'of', 'Uniqueness:', 'Extending', 'Valiant-Vazirani', 'Theorem', 'to', 'the', 'Probabilistic', 'and', 'Quantum', 'Settings']	['The', 'Pursuit', 'of', 'Uniqueness', 'Extending', 'ValiantVazirani', 'Theorem', 'to', 'the', 'Probabilistic', 'and', 'Quantum', 'Settings']	['Pursuit', 'Uniqueness', 'Extending', 'ValiantVazirani', 'Theorem', 'Probabilistic', 'Quantum', 'Settings']	['pursuit', 'uniqueness', 'extending', 'valiantvazirani', 'theorem', 'probabilistic', 'quantum', 'settings']

4.3 Feature Extraction

Feature extraction was conducted using two methods: TFIDF and SBERT, with the results presented in Table 4. The TFIDF method generated a feature vector with dimensions of 17.762, reflecting the size of the vocabulary. However, this vector was sparse, with many values being zero, indicating that only a subset of the vocabulary terms were present in the input text. In contrast, the SBERT method produced a much more compact feature vector with dimensions of 768, where all elements were dense, meaning every part of the vector was filled with meaningful information. This difference highlights how TFIDF captures term frequency and importance within a large vocabulary, often resulting in sparse representations, whereas SBERT creates a more condensed, context-aware representation that fully utilizes the vector space.

Table 4. Results of TFIDF and SBERT feature extraction

Input text	TFIDF	SBERT
['pursuit', 'uniqueness', 'extending', 'valiantvazirani', 'theorem', 'probabilistic', 'quantum', 'settings']	[0, 0, 0, 0, 0, 0, 0, …, 0, 0, 0, 0, 0, 0, 0]	[-3.51481279e-03, 4.55926582e-02 -1.39396433e-02, -1.08676273e-02, …, -5.71076851e-03, -2.77706049e-03, -6.65656291e-03, 3.07868775e-02]

4.4 Modelling

The modeling phase of this study involves leveraging Convolutional Neural Networks (CNN) for dimensionality reduction, followed by feeding the processed data into an ensemble learning framework. The proposed CNN architecture, illustrated in Figure 4, begins with reshaping the 768-dimensional SBERT output into a 28x28x1 format through zero padding for the remaining dimensions. This transformation prepares the data for efficient processing within the CNN layers. The reshaped data is then passed through two convolutional layers, each utilizing the Exponential Linear Unit (ELU) activation function. This function is selected for its ability to improve the learning

dynamics and overall performance of deep neural networks by mitigating the vanishing gradient problem. Following the convolutional layers, a flatten layer is applied to convert the multi-dimensional data into a one-dimensional vector, which is then fed into a fully connected layer with an output size of 512. The resulting 512-dimensional output, termed the latent feature vector, encapsulates the essential features extracted from the original high-dimensional SBERT embeddings. This integration of SBERT and CNN allows the model to combine rich, context-aware embeddings with CNN's ability to capture local and hierarchical patterns, thereby enhancing the quality of feature representations. CNN also performs non-linear dimensionality reduction, enabling the model to emphasize the most salient aspects of the input, reducing noise and redundancy. This latent feature vector is subsequently fed into the ensemble learning model for training, aiming to enhance the classification performance by leveraging the strengths of both CNN and ensemble methods.

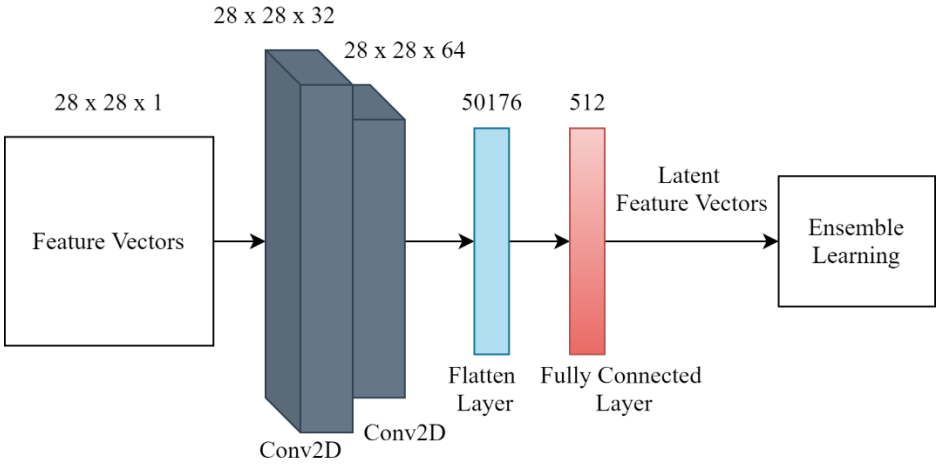

Fig. 4. Proposed CNN architecture

To ensure that this compression does not result in significant information loss, a reconstruction architecture using a decoder (mirroring the encoder's structure) was implemented. This encoder-decoder (autoencoder) setup was trained using the Adam optimizer with a learning rate of 1e-3 for 50 epochs. The fidelity of the compressed representation was evaluated using the MSE between the original SBERT embeddings and the reconstructed outputs. The resulting MSE value of 0.00127 indicates that the dimensionality reduction process effectively preserved the essential information from the original feature space, with minimal loss. This confirms that the proposed CNN architecture successfully captures meaningful latent features while reducing dimensionality in a non-linear manner. The training loss over the 50 epochs is illustrated in Figure 5, which shows a consistent decline, indicating stable convergence of the model during the reconstruction process.

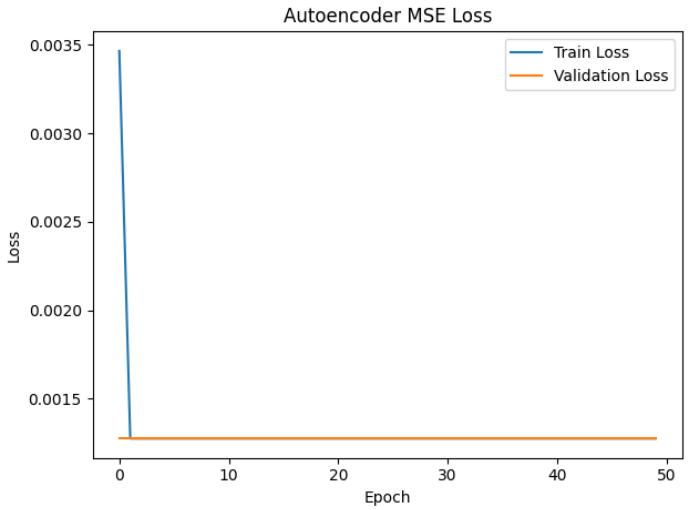

Fig. 5. Training loss curve of the CNN encoder-decoder architecture

4.5 Evaluation

The model evaluation process employed precision, recall, F1 score, and accuracy as key metrics to assess performance comprehensively. Evaluations were conducted using features derived from both title-only inputs and combined title and abstract inputs. Feature extraction methods utilized in this analysis included TFIDF, SBERT, and a combination of SBERT with CNN. By applying these metrics across different types of input data and extraction techniques, the evaluation aimed to provide a detailed understanding of how each model performed, highlighting the impact of both input complexity and feature representation on overall model effectiveness. The results of all experiments are presented in Tables 5-7.

Table 5. Model performance using TFIDF feature extraction

No	Model	Precision (%)		Recall (%)		F1-Score (%)		Accuracy (%)	
		Title	Title + Abstract	Title	Title + Abstract	Title	Title + Abstract	Title	Title + Abstract
1	Decision Tree	80,39	84,55	80,59	84,84	80,45	84,67	80,59	84,84
2	Adaboost	74,15	87,75	72,41	87,93	70,69	87,81	72,41	87,93
3	Random Forest	86,29	90,37	86,38	89,46	85,76	88,31	86,38	89,46
4	XGBoost	85,11	**93,35**	85,11	**93,47**	84,68	**93,36**	85,11	**93,47**

The experiment results demonstrated in table 5 that incorporating abstracts into the input data for ensemble learning models significantly enhances classification accuracy. Notably, Adaboost's performance saw a substantial improvement with the addition of abstracts with accuracy rising from 72.41% to 87.93%, but XGBoost still outperformed all other models, emerging as the best performer. This underscores the value of richer textual information in boosting model effectiveness, with XGBoost leading the way in harnessing this added context.

Table 6. Model performance using SBERT feature extraction

No	Model	Precision (%)		Recall (%)		F1-Score (%)		Accuracy (%)	
		Title	Title + Abstract	Title	Title + Abstract	Title	Title + Abstract	Title	Title + Abstract
1	Decision Tree	74,12	85,35	74,01	85,40	74,06	85,37	74,01	85,40
2	Adaboost	79,60	88,32	80,44	88,52	79,49	88,33	80,44	88,52
3	Random Forest	86,32	**93,67**	86,40	**93,67**	85,29	**93,37**	86,40	**93,67**
4	XGBoost	87,61	88,54	87,97	88,84	87,63	88,12	87,97	88,84

The experiment revealed in table 6 that Decision Tree showed a notable improvement, with accuracy rising from 74.01% to 85.40%. Adaboost also benefited, with accuracy increasing from 80.44% to 88.52%. Random Forest experienced the most substantial gain, with accuracy improving from 86.40% to 93.67%, making it the top performer. Random Forest demonstrating the greatest effectiveness in utilizing the added context when the feature is extracted using SBERT.

Table 7. Model performance using SBERT feature extraction on hybrid model

No	Model	Precision (%)		Recall (%)		F1-Score (%)		Accuracy (%)	
		Title	Title + Abstract	Title	Title + Abstract	Title	Title + Abstract	Title	Title + Abstract
1	CNN + Decision Tree	72,45	85,05	72,37	84,94	72,41	84,99	72,37	84,94
2	CNN + Adaboost	79,16	89,12	79,94	89,26	79,24	89,15	79,94	89,26
3	CNN + Random Forest	86,71	93,75	86,83	93,71	85,78	93,38	86,83	93,71
4	CNN + XGBoost	88,81	**94,51**	89,08	**94,62**	88,86	**94,51**	89,08	**94,62**

The experiment revealed in table 7 that incorporating abstracts significantly boosts classification accuracy across models using SBERT embeddings and CNN for features dimension reduction. Adaboost showed a notable improvement with the addition of abstracts, reflecting its sensitivity to richer context. However, XGBoost, while also benefiting from the abstracts, emerged as the best performer overall, demonstrating superior effectiveness in leveraging the enhanced features provided by CNN and SBERT. Subsequently, the answers to research questions Q1, Q2, and Q3 will be addressed.

1) Answer for Q1:

Figure 6 illustrates the performance of models when evaluated using different input scenarios: title-only versus the combined title and abstract. This comparison highlights how the inclusion of abstracts impacts model effectiveness across various feature extraction techniques and models, providing insights into the added value of richer textual information in enhancing classification accuracy and other evaluation metrics.

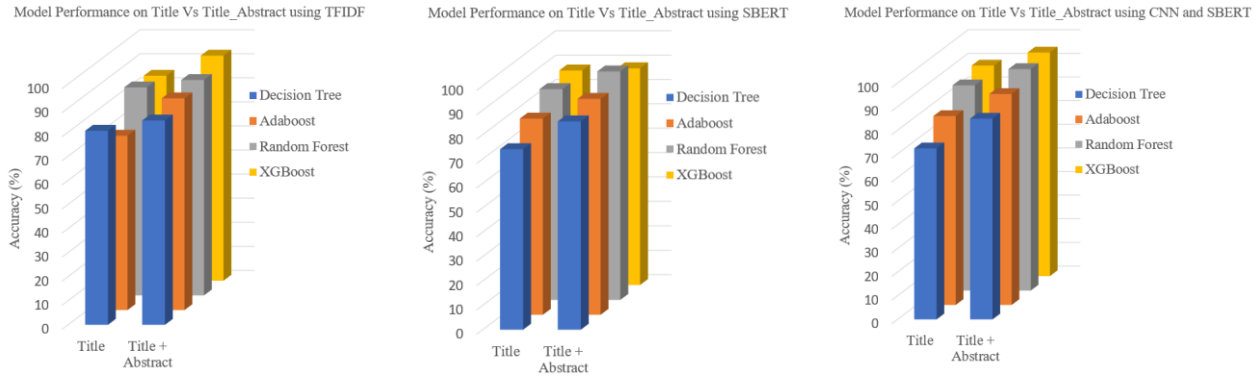

Fig. 6. Model performance on title versus title_abstract in all schenario

The experiments results in figure 6 demonstrated that incorporating abstracts significantly enhances classification accuracy across various models, with different techniques showing distinct advantages. Using TFIDF, Adaboost's performance improved notably with the addition of abstracts, though XGBoost remained the top performer overall. With SBERT embeddings, the integration of abstracts led to substantial gains in accuracy for all models, with Random Forest achieving the highest performance. In the CNN + SBERT scenario, Adaboost benefited from the richer context provided by the abstracts, but XGBoost excelled as the best performer, leveraging the enhanced features most effectively. These findings highlight the critical role of richer textual information in boosting model effectiveness, with XGBoost emerging as the most effective model in harnessing these enhancements.

2) Answer for Q2:

Figure 7 presents a comparison of model performance when using the combined title and abstract as input features, extracted with two different methods: TFIDF and SBERT. The previous experiments demonstrated that incorporating both the title and abstract improves performance due to the richer textual information available. This figure highlights the impact of each feature extraction method—TFIDF and SBERT—on the models' effectiveness, showing how different approaches to handling the combined text influence overall performance.

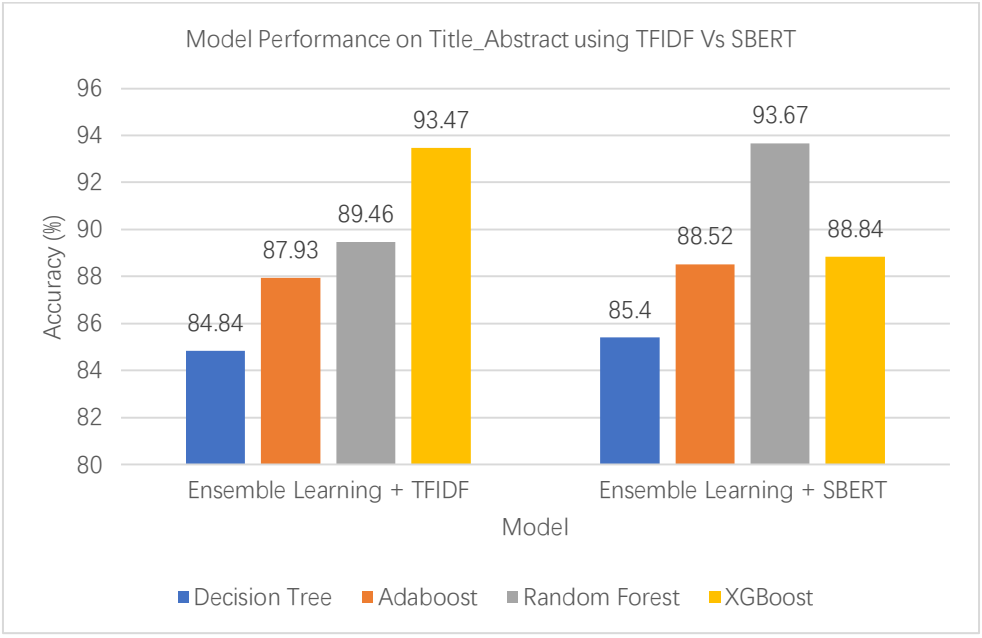

Fig. 7. Model performance on title_abstract using TFIDF vs SBERT

The analysis reveals of figure 7 that Decision Tree, Adaboost, and Random Forest all experienced improvements with SBERT embeddings, reflecting the benefits of richer contextual information. Random Forest achieved the highest accuracy with SBERT-generated features, demonstrating its strong capability to leverage advanced embeddings effectively. However, XGBoost showed a decrease in performance with SBERT, suggesting that the enhanced features did not align well with its model architecture in this case. Despite the advantages of SBERT in providing contextual embeddings, the TFIDF method, a conventional approach that does not leverage contextual information, still managed to closely rival SBERT's performance. For instance, XGBoost performed better with TFIDF than with SBERT, and

Adaboost also showed competitive results with TFIDF. This indicates that traditional feature extraction methods can be highly effective and competitive, even compared to advanced, context-aware techniques. Overall, while SBERT provides significant improvements in many scenarios, TFIDF remains a strong contender, underscoring that both conventional and advanced methods have their place depending on the context and model used. Furthermore, the difference in model performance can be attributed to the nature of the feature representations. TFIDF produces sparse, high-dimensional vectors that tend to work well with models like XGBoost, which are optimized for handling sparse input [27]. TFIDF is purely based on word frequency and does not capture semantic relationships or contextual meaning within the text, making it more effective for models that rely on discrete and independent features. In contrast, SBERT generates dense, low-dimensional embeddings that capture semantic context, making them more suitable for models like Decision Tree, AdaBoost, and Random Forest. The mismatch between SBERT's dense features and XGBoost's architecture may explain the observed performance drop.

3) Answer for Q3:

Figure 8 presents a comparison of model performance using the combined title and abstract as input features, with SBERT and CNN + SBERT as feature extraction methods. This figure highlights the effectiveness of integrating CNN with SBERT, demonstrating how combining CNN's feature dimension reduction with SBERT's contextual embeddings enhances model performance. Despite the reduction of feature dimensions through CNN, which generates latent feature vectors, CNN retains the rich contextual information provided by the text. This comparison illustrates the impact of each approach on model performance, emphasizing that CNN + SBERT maintains the depth of contextual information while optimizing feature representation.

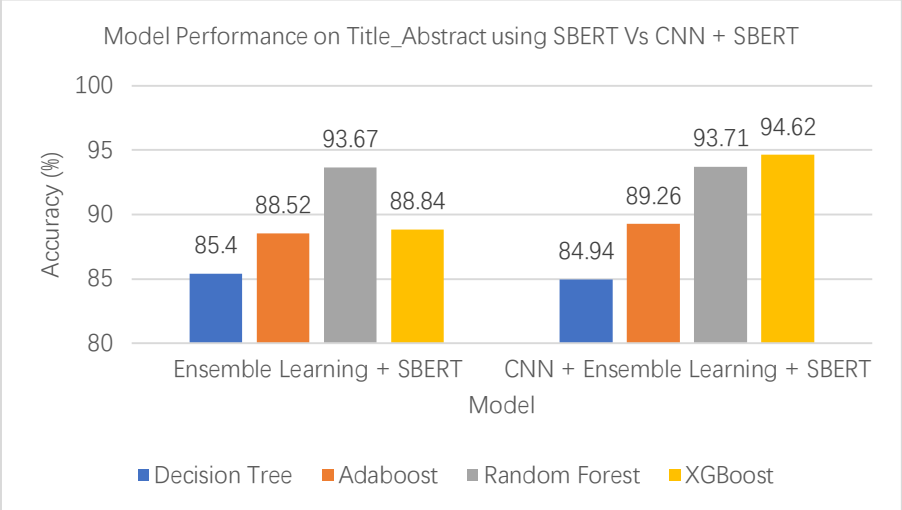

Fig. 8. Model performance on title_abstract using SBERT vs CNN + SBERT

Based on figure 8, the analysis shows that integrating CNN with SBERT resulted in performance improvements for all models except Decision Tree, which experienced a slight decrease. This indicates that CNN's feature reduction may not have benefited Decision Tree as much. CNN + SBERT fed into Adsaboost achieved an accuracy of 89.26%, surpassing its performance with SBERT alone, while CNN + Random Forest reached 93.71%, demonstrating strong results comparable to its performance with SBERT. CNN + SBERT using XGBoost attained the highest accuracy of 94.62%, significantly outperforming XGBoost with SBERT alone. Ensemble models like XGBoost and Random Forest are known for their excellent performance in handling complex and diverse features, while also reducing overfitting through the aggregation of multiple decision trees [28]. Moreover, their performance is often comparable, with no statistically significant differences reported in various classification tasks [29,30]. CNNs, though commonly used in image tasks, effectively capture hierarchical patterns and have been successfully applied to dense representations like SBERT embeddings. The latent features generated by CNNs are highly suitable for ensemble models like Random Forest and XGBoost, allowing them to focus on high-level features that are most relevant for accurate predictions. Overall, the combination of CNN with SBERT enhanced model performance across most models, highlighting the potential for further improvements in performance through effective feature reduction. This approach demonstrates that CNN+SBERT generally provides a boost, optimizing feature reduction strategies could offer additional gains in performance.

Consistent with prior research emphasizing the benefits of hybrid models [31] our proposed approach, which combines CNN and ensemble learning, reinforces these findings by demonstrating substantial performance gains over standalone ensemble models. The incorporation of CNN in this hybrid model effectively captures deeper contextual features from the data that traditional ensemble models may overlook. By combining CNN's spatial feature extraction capabilities with the generalization strengths of ensemble learning, the model achieves greater robustness and accuracy

in classification tasks. The findings from the literature [1]**,** which suggest that increasing the amount of text studied can improve classification evaluation, our research uses relevant attributes to enhance the model's ability to classify text more accurately. The comparison between BERT and CNN+BERT will provide deeper insights into the impact of dimensionality reduction on classification performance, as well as the potential of this combination to produce optimal feature representations for text classification tasks. CNNs are particularly effective in scenarios where detecting local and position-invariant patterns is essential, such as identifying key phrases or specific topics within text. This characteristic makes CNNs one of the most popular architectures for text classification. In our research, the use of CNN as part of a hybrid model with ensemble learning aligns with this strength. The ability of CNN to capture important local features contributes significantly to the model's overall performance [32]**.**

5. Conclusion

The research demonstrates that incorporating both the title and abstract significantly enhances classification accuracy across various models, underscoring the value of richer textual data in improving model performance. Feature extraction methods like SBERT and its combination with CNN lead to dense, contextually rich embeddings that generally outperform traditional TFIDF vectors, which are often sparse and less informative. Among the models tested, CNN + SBERT with XGBoost achieved the highest accuracy of 94.62%, indicating that integrating advanced feature extraction techniques with powerful models can yield superior results. However, this study relies on a single dataset (arXiv), which may limit the generalizability of the results. For future research, the proposed model will be tested on multiple datasets to evaluate its generalizability and scalability. The findings also suggest that while CNN + SBERT provides a notable boost, further optimization of feature reduction strategies could offer additional performance improvements. Additionally, exploring the use of sequential models, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, could be a promising avenue for future research, potentially enhancing the ability to capture and leverage temporal dependencies in text data for even better model performance.

References

[1] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Eng. J.*, vol. 60, no. 3, pp. 3401–3409, 2021, doi: 10.1016/j.aej.2021.02.009.

[2] Q. Li *et al.*, "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, 2022, doi: 10.1145/3495162.

[3] M. Rivest, E. Vignola-Gagné, and É. Archambault, "Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling," *PLoS One*, vol. 16, no. 5 May, pp. 1–18, 2021, doi: 10.1371/journal.pone.0251493.

[4] M. Gao *et al.*, "A Novel Hierarchical Discourse Model for Scientific Article and It's Efficient Top-K Resampling-based Text Classification Approach," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 2022-Octob, no. 2, pp. 774–781, 2022, doi: 10.1109/SMC53654.2022.9945306.

[5] S. Zhang, S. Wang, R. Liu, H. Dong, X. Zhang, and X. Tai, "A bibliometric analysis of research trends of artificial intelligence in the treatment of autistic spectrum disorders," *Front. Psychiatry*, vol. 13, 2022, doi: 10.3389/fpsyt.2022.967074.

[6] S. Chowdhury and M. P. Schoen, "Research Paper Classification using Supervised Machine Learning Techniques," *2020 Intermt. Eng. Technol. Comput. IETC 2020*, 2020, doi: 10.1109/IETC47856.2020.9249211.

[7] M. Miric, N. Jia, and K. G. Huang, "Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents," *Strateg. Manag. J.*, vol. 44, no. 2, pp. 491–519, 2023, doi: 10.1002/smj.3441.

[8] A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, "A systematic review of text classification research based on deep learning models in Arabic language," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 6, pp. 6629–6643, 2020, doi: 10.11591/IJECE.V10I6.PP6629-6643.

[9] D. N. Reddy, S. M. Bagali, and S. Sadiya, "Machine Learning Algorithms for Detection: A Survey and Classification," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 10, pp. 3468–3475, 2021.

[10] H. Manoharan *et al.*, "A machine learning algorithm for classification of mental tasks," *Comput. Electr. Eng.*, vol. 99, no. February, p. 107785, 2022, doi: 10.1016/j.compeleceng.2022.107785.

[11] M. Canaparo, S. C. Todeschini, and E. Ronchieri, "Transformer-based Architecture for Assisting Title and Abstract Screening of a Systematic Review," pp. 1–1, 2023, doi: 10.1109/nssmicrtsd49126.2023.10338654.

[12] I. N. Switrayana and N. U. Maulidevi, "Collaborative Convolutional Autoencoder for Scientific Article Recommendation," *Proc. - 2022 9th Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2022*, pp. 96–101, 2022, doi: 10.1109/ICITACEE55701.2022.9924130.

[13] P. L. Teh and C. F. Uwasomba, "Impact of Large Language Models on Scholarly Publication Titles and Abstracts: A Comparative Analysis," *J. Soc. Comput.*, vol. 5, no. 2, pp. 105–121, 2024, doi: 10.23919/JSC.2024.0011.

[14] J. Eykens, R. Guns, and T. C. E. Engels, "Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches," *Quant. Sci. Stud.*, vol. 2, no. 1, pp. 89–110, 2021, doi: 10.1162/qss_a_00106.

[15] J. Dai and C. Chen, "Text classification system of academic papers based on hybrid Bert-BiGRU model," *Proc. - 2020 12th Int. Conf. Intell. Human-Machine Syst. Cybern. IHMSC 2020*, vol. 2, pp. 40–44, 2020, doi: 10.1109/IHMSC49165.2020.10088.

[16] M. M. Ahanger and M. A. Wani, "Novel Deep Learning Approach for Scientific Literature Classification," *Proc. 2022 9th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2022*, pp. 249–254, 2022, doi: 10.23919/INDIACom54597.2022.9763218.

[17] M. M. Ahanger and M. A. Wani, "Comparative analysis of deep learning approaches for scientific literature classification," *Proc. 2021 8th Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2021*, pp. 74–80, 2021, doi: 10.1109/INDIACom51348.2021.00015.
[18] J. Zhang, K. Chusap, W. Zhang, and C. Liu, "Improving Paper Classification Using Forecasting," *Proc. - 2022 IEEE Int. Conf. Big Data, Big Data 2022*, pp. 2454–2460, 2022, doi: 10.1109/BigData55660.2022.10020764.
[19] C. Fan, Y. Li, and Y. Wu, "Multi feature fusion paper classification model based on attention mechanism," *Proc. - 2023 5th Int. Conf. Nat. Lang. Process. ICNLP 2023*, pp. 308–312, 2023, doi: 10.1109/ICNLP58431.2023.00063.
[20] A. Mulahuwaish, K. Gyorick, K. Z. Ghafoor, H. S. Maghdid, and D. B. Rawat, "Efficient classification model of web news documents using machine learning algorithms for accurate information," *Comput. Secur.*, vol. 98, 2020, doi: 10.1016/j.cose.2020.102006.
[21] I. M. Rabbimov and S. S. Kobilov, "Multi-Class Text Classification of Uzbek News Articles using Machine Learning," *J. Phys. Conf. Ser.*, vol. 1546, no. 1, pp. 0–11, 2020, doi: 10.1088/1742-6596/1546/1/012097.
[22] N. Sulistianingsih and I. N. Switrayana, "Enhancing Sentiment Analysis for the 2024 Indonesia Election Using SMOTE-Tomek Links and Binary Logistic Regression," *Int. J. Educ. Manag. Eng.*, vol. 14, no. 3, pp. 22–32, 2024.
[23] B. Devlin and R. Liu, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.
[24] H. Choi, J. Kim, S. Joe, and Y. Gwon, "Evaluation of BERT and Albert sentence embedding performance on downstream NLP tasks," *Proc. - Int. Conf. Pattern Recognit.*, pp. 5482–5487, 2020, doi: 10.1109/ICPR48806.2021.9412102.
[25] I. N. Switrayana, R. Hammad, P. Irfan, T. T. Sujaka, and M. H. Nasri, "Comparative Analysis of Stock Price Prediction Using Deep Learning with Data Scaling Method," vol. 7, no. 1, pp. 78–90, 2025.
[26] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Comparative Analysis using Various Performance Metrics in Imbalanced Data for Multi-class Text Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 6, pp. 1082–1090, 2023, doi: 10.14569/IJACSA.2023.01406116.
[27] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-August-2016, pp. 785–794, 2016, doi: 10.1145/2939672.2939785.
[28] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies*, vol. 13, no. 3, pp. 1–40, 2025, doi: 10.3390/technologies13030088.
[29] Y. Ibrahim, E. Okafor, B. Yahaya, S. M. Yusuf, Z. M. Abubakar, and U. Y. Bagaye, "Comparative Study of Ensemble Learning Techniques for Text Classification," *2021 1st Int. Conf. Multidiscip. Eng. Appl. Sci. ICMEAS 2021*, no. May, 2021, doi: 10.1109/ICMEAS52683.2021.9692306.
[30] Z. Shao, M. N. Ahmad, and A. Javed, "Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface," *Remote Sens.*, vol. 16, no. 4, 2024, doi: 10.3390/rs16040665.
[31] S. Gonçalves, P. Cortez, and S. Moro, "A deep learning classifier for sentence classification in biomedical and computer science abstracts," *Neural Comput. Appl.*, vol. 32, no. 11, pp. 6793–6807, 2020, doi: 10.1007/s00521-019-04334-2.
[32] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep Learning-Based Text Classification: A Comprehensive Review," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, 2021, doi: 10.1145/3439726.

Authors' Profiles

I. Nyoman Switrayana, currently work as a Lecturer in the Department of Computer Science at Bumigora University. He received his bachelor's degree in Computer Engineering (S.T) from Telkom University, Bandung. His master's degree (M.T) was obtained from the School of Electrical Engineering and Informatics, Bandung Institute of Technology. His area of interest includes Artificial Intelligence, Natural Language Processing, Computer Vision, Data Mining, Machine Learning, Deep Learning, and Recommender Systems.

Neny Sulistianingsih obtained a bachelor's degree in informatics engineering (S.Kom) and a master's degree in Informatics Engineering (M.Kom) from the Universitas Islam Indonesia, Yogyakarta. She has just completed his doctoral education at Universitas Gadjah Mada, Yogyakarta. Her research interests include data mining, sentiment analysis, big data, text mining, graph mining, and machine learning. She is currently a lecturer who teaches at Universitas Bumigora, Mataram.

How to cite this paper: I. Nyoman Switrayana, Neny Sulistianingsih, "Leveraging Convolutional Neural Network to Enhance the Performance of Ensemble Learning in Scientific Article Classification", International Journal of Modern Education and Computer Science(IJMECS), Vol.17, No.6, pp. 146-159, 2025. DOI:10.5815/ijmecs.2025.06.10