

A Hybrid Deep Learning Framework for Scientific Correction Text Classification

Garima Sharma*, Vikas Tripathi, Vijay Singh

Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun, Uttarakhand, India

Abstract—Ensuring research integrity and maintaining the reliability of scientific communication requires a clear and comprehensive understanding of the underlying causes of article correction reasons across domains. Correction notices issued to authors are often unstructured with inconsistent textual information, making their automatic classification a challenging task. Existing approaches primarily rely on manual analysis or conventional machine learning techniques that fail to capture the contextual semantics and complex linguistic patterns present in the given correction statements. To address this limitation, this study proposes a deep learning-based framework for the automated classification of correction reasons using a large-scale dataset. The approach integrates contextual semantic representations with a hybrid classification strategy that combines data normalization, ensemble learning, and soft voting mechanisms to enhance classification performance and overall predictive stability. The experimental results demonstrate that the framework outperforms conventional machine learning and existing deep learning algorithms across multiple evaluation metrics. The model achieves an accuracy of 94.3%, along with strong specificity and predictive reliability. The model consistently demonstrates strong evaluation metrics, such as precision, recall, and F1-score, showcasing its effectiveness in handling a complex and imbalanced retraction dataset. The experimental analysis was conducted on a dataset categorized into four distinct classes, namely Correction, Expression of Concern, Reinstatement, and Retraction. The findings of the study offer a scalable and robust solution for systematically identifying correction reasons and uncovering underlying patterns. By enabling transparent, consistent, and efficient interpretation of correction information, this proposed approach can support publishers, academic institutions, policymakers, and researchers in detecting the patterns of scientific misconduct and thus facilitating informed decision-making while mitigating the risk of misinformation in scientific communication.

Keywords—Correction analysis; research integrity; BERT; natural language processing; text classification

I. INTRODUCTION

Scholarly publications guarantee the reliability of scientific research. Ensuring the accuracy and credibility of their findings is critical due to their direct and exponential impact on the community health sector [1]. Journals and Publishers employ several mechanisms such as issuing errata for minor errors, expressing concerns, and, in more severe cases, retracting the article to maintain the integrity of the scientific record [2]. A retraction serves to alert readers to articles with significant flaws or erroneous data, rendering their conclusions and findings untrustworthy [3]. These can occur for a variety of reasons, such as fraud, methodological errors, data set and results issues,

plagiarism, and publisher errors, as indicated [4]. Although sparsely observed, the rate of withdrawals has increased in recent decades. Even after being retracted, these articles continue to be cited and included in meta-analyses, as evidenced [5], potentially compromising the integrity of the findings in the cited articles. However, no studies have specifically investigated withdrawals in the literature [6]. Authors in [7] aimed to explore the frequency and features of the withdrawn articles investigated and to evaluate the extent to which these retracted publications are cited. Thus, the author who submits manuscripts has the responsibility to ensure that references in their bibliography are not retracted. If an author (or reader) notices that a published article cites a retracted article, they should take the initiative to notify the journal for a correction or withdrawal [8]. The rate and magnitude of withdrawals serve as key indicators of overall operations within the domain of scientific and technical publications, highlighting the need for detailed studies [9]. In recent years, withdrawals have become a significant area of research both nationally and internationally [10]. Three key trends that have emerged in withdrawal patterns are: 1) the rate of retracted published research is growing rapidly, supported by substantial evidence [11]; 2) withdrawals are no longer limited to researchers from developed countries, but have also been extended to developing countries across academic disciplines [12]; and 3) factors of manuscript withdrawal can be grouped into two broad categories: data misconduct and unintentional or honest errors [13].

The integrity of a scholarly publication requires an effective mechanism to detect and remove falsified publications. Recently, there has been a significant increase in the withdrawal of published articles, bringing immediate attention to the reliability of publication practices [14]. These withdrawals support an essential corrective process to prevent the dissemination of false and misleading findings that adversely impact decision-making and scientific progress has been underpinned [15]. In recent years, the reasons for any withdrawal of an article have been too diverse, requiring a systematic classification mechanism. It has been observed that a large body of related research has been carried out using publicly available datasets such as Retraction Watch, which comprise a wide range of factors used to identify an illegitimate publisher [16]. However, with a larger consolidation, the primary reasons are data inconsistency, ethical misconduct, publication malpractices, and erroneous methodologies present, as well as the associated results [17]. Inconsistent, unreliable, and missing data-related information has been found to be dominant in data-related concerns has been observed [18]. In contrast, plagiarism, fabrication, falsification, and, in fact, peer review manipulations

*Corresponding author.

are related to ethical misconduct that undermines scientific credibility [19]. These concerns are widely reported and frequently observed across sensitive domains such as cardiovascular and clinical research, etc., increasing with time has been demonstrated [20].

Thus, the withdrawal of the articles represents a pivotal part of a self-correcting mechanism of science that supports the removal of unreliable, invalid, and ethically compromised findings from scholarly records. This plays a significant role in ensuring academic integrity and transparency of scientific communication in an evidence-based healthcare system. Twitter data to identify issues with publications. They focused on analyzing tweet texts related to three withdrawals of COVID-19 and SARS-CoV-2 publications and their withdrawal notices [21]. Although no early warning signs were found in tweets for one publication, tweets about the other two raised concerns about their validity shortly after publication. The experiments showcased that Twitter could play a role in identifying questionable publications. In 2022, [22] aimed to check how well major search engines, such as Google Scholar, and scientific ones such as PubMed, EMBASE, SCOPUS, and Web of Science, could detect article withdrawal. The retracted articles were taken from the Retraction Watch Database. Although Google Scholar retrieved the majority of the articles, the results revealed that it was not as reliable as scientific databases in identifying the retracted articles. The study recommended using different scientific search engines instead of just a general one. This would help avoid using retracted articles and ensure more reliable research.

To analyze the rate and character of withdrawal articles comprising the literature on infectious diseases and clinical microbiology, [23] conducted a study in which they used the Retraction Watch Database for a cross-sectional analysis of publications retracted up to 30 June 2022 in the said fields. In addition, articles with available citation data from the Web of Science database were also included in the study. The work explored trends in withdrawals and the characteristics of these publications. It also examined the number of references to the articles before and after the withdrawal year. [24] examined the transparency and consistency of the way public health publications were presented to researchers. It was a cross-sectional study that included 441 articles in public health. Data were collated from 11 sources for each publication, with withdrawal notices obtained from publisher websites and full-text aggregators. The criteria of the National Library of Medicine and the Committee on Publication Ethics were used to determine whether an article had been withdrawn. In another work presented [25], data are analyzed on retracted articles to identify trends in withdrawals. Publications were searched in WoSCC and imported into CiteSpace in plain text format for analysis. Through a visual examination of trends in countries, institutions, keywords, and subject areas, the patterns surrounding the retracted articles were revealed. The analysis also highlighted the fields most affected by withdrawals. The comparative overview of recent related work on the classification of withdrawals, highlighting methodological approaches, key findings, and research gaps, is given in Table V (in the Appendix). The existing literature shows significant progress in the identification and categorization of causes for

withdrawals. There exists a gap in a universal approach and a computationally robust classification framework. For improving transparency and large-scale analysis, a robust automated system can help address these identified gaps and maintain the integrity of scientific knowledge. In addition, integrative approaches capable of combining domain expertise with data-driven methodological frameworks can help to develop more standardized and scalable classification models. The advancement in artificial intelligence (AI) and its subdomains has recently made it possible to design and develop automated systems that can detect the withdrawal reason of withdrawn articles with high precision and efficiency, as highlighted in [35]. These systems work by analyzing patterns in metadata, textual content, and article history to determine if an article has been retracted and categorize it accordingly [36]. The problem with using retracted literature is that its inclusion in literature reviews can mislead future research and clinical practice. To address these issues, a multi-representation semantic learning framework integrating an enhanced Word2Vec embedding with GloVe and Sentence-BERT to capture lexical, semantic, and contextual information from scientific withdrawal notices has been proposed in this work. These complementary representations are subsequently classified as a hybrid deep learning-based AAHA-ShLNet architecture, offering three key contributions:

- A multi-representation feature extraction framework using Enhanced Word2Vec, GloVe, and Sentence-BERT for scientific withdrawal text.
- Development of a hybrid AAHA-ShLNet classifier integrating AAHA-BN-ShNet, LeNet, and soft voting for robust classification.
- Validation of the proposed framework on 52,811 scientific post-publication notice records using a group-aware partitioning mechanism, demonstrating its effectiveness across multiple evaluation metrics.

II. MATERIALS AND METHODS

The work proposes a hybrid deep learning framework to classify withdrawal-related categories. The framework, as given in Fig. 1, integrates an advanced text representation technique along with a mutually reinforcing neural network architecture to improve the accuracy and robustness of the classification framework. The comprehensive workflow consists of three main sequential phases: 1) Accumulated Data Preprocessing, 2) Key Feature Engineering, and 3) Hybrid Classification.

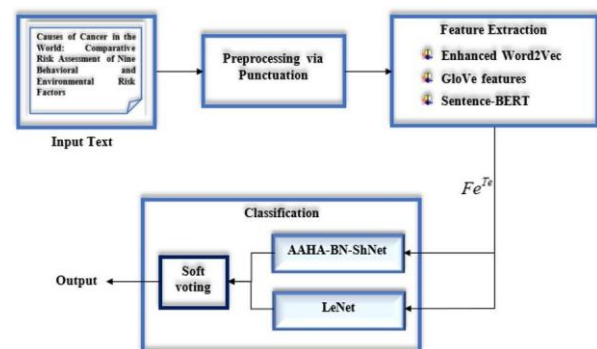


Fig. 1. Workflow architecture of proposed framework

A. Data Preprocessing via Punctuation

1) *Dataset and class distribution*: The proposed work utilises the four-class post-publication retraction notice records obtained from the Retraction Watch Database for four-class post-publication notice classification. The target variable, Retraction Nature, comprises four editorial-action categories: Correction, Expression of Concern, Reinstatement, and Retraction; the description of each is given in Table. I. After loading the database using an encoding-compatible procedure and validating the record structure, the resulting corpus contained 52,811 records.

TABLE I. DESCRIPTION OF POST PUBLICATION NOTICES

Category	Description
Correction	Published article requires correction, without constituting a full retraction of the article.
Expression of Concern	Concern regarding the reliability, integrity, or validity of a publication.
Reinstatement	Retracted publication has been restored to the active literature databases.
Retraction	Withdrawal of a published article from the scientific record

2) *Data preprocessing*: Text Processing is used to standardize and clean the collated input data, Tei, prior to feeding the same to the feature extraction engine. Specifically, punctuation has been handled along with the removal of redundant symbols without disturbing meaningful delimiters such as periods and commas. This step also ensures that the sentence boundaries and contextual structure remain intact while reducing the overall noise in the refined dataset. In addition, the output of this stage is punctuated text, TPej, serving as input for the subsequent extraction of features.

B. Feature Extraction

To capture the syntactical and semantic characteristics of the text effectively, a multi-representation feature extraction strategy has been adopted comprising three complementary embedding techniques:

- **Enhanced Word2Vec**: For dense vector representations of words using CBOW and Skip-Gram architectures as shown in Fig. 2, Word2Vec is utilized. Word2Vec is a shallow neural network model applied to learn word embeddings in NLP.

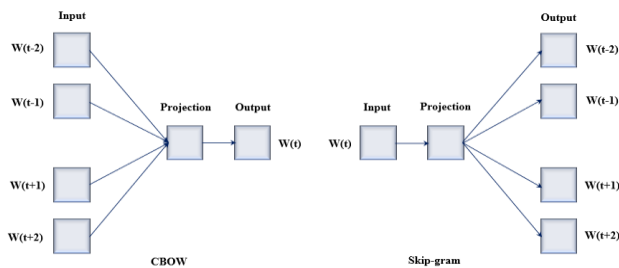


Fig. 2. Word2Vec architectural diagram.

The standard cosine similarity measures the angular relationship between two vectors, keeping the embedding

dimensions equal. In scientific publication withdrawal texts, embeddings are often sparse, noisy, and have a highly imbalanced distribution due to domain-specific terminologies and unequal class frequencies. To improve the capability of the used architectures and capture semantic similarity within complex text more effectively, a modified measure of cosine similarity is introduced. This modification incorporates the following:

- An adaptive scaling factor to adjust similarity based on data variability and to reduce the influence of extreme embedding values.
- Log-exp transformation to reduce the influence of extreme values and stabilize semantic distances. This enhancement improves the robustness of the system to capture the semantic relationships across diverse textual distributions. The resulting embeddings are denoted as:

$$\cos(\theta) = \{X \cdot Y\} / \{|X||Y|\} \quad (1)$$

The most significant similarity measure utilized by Word2Vec for word vector comparison is cosine similarity, which measures the angle between two vectors as defined by Eq. (1). Increasing the cosine similarity score (towards 1) indicates an increase in similarity among the words or vectors. This measure of similarity can be very important in classifying articles on the basis of some features. However, Word2Vec does not capture the entire similarity between whole sentences or documents. Word2Vec operates by evaluating similarities among word vectors, but when combining these vectors for sentence or document understanding, the process becomes limited. In order to overcome these issues, a modification to the standard cosine similarity measure used by Word2Vec. An adaptive scaling factor δ is used, which adjusts the similarity measure based on the variability of the dataset. This allows the model to adapt to different types of text distributions, making it more robust and accurate, especially in complex text domains. Using the log-exp transformation, the extreme values are smoothed, resulting in a similarity measure that is more balanced and less sensitive to outliers. The updated version of the cosine similarity equation [37], as given and utilized, is defined in Eq. (2).

$$MS(X, Y) = -\log(1 - \exp(-\delta \cdot \cos(X, Y))) \quad (2)$$

To compute cosine similarity between two vectors X and Y , the dot product is divided by the product of the vectors' magnitudes [38], as given in Eq. (3).

$$\cosine(X, Y) = \sum_{i=1}^n X_i * Y_i / \sqrt{\sum_{i=1}^n X_i^2} * \sqrt{\sum_{i=1}^n Y_i^2} \quad (3)$$

The scaling parameter δ can be calculated using Eq. (4).

$$\delta = (\cos(X, Y) - \min(\cos(X, Y)) / \text{mean}(\cos(X, Y))) \quad (4)$$

The four-step process for computing the modified cosine similarity includes:

1) Compute the modified cosine similarity based on the updated expression in Eq. (2).

2) The distance (d) between two vectors X and Y is obtained using Eq. (5).

$$d = 1 - \text{ModifiedSimilarity}(X, Y) \quad (5)$$

3) Depending on the distance value d , if $d < 0.5$, then the returned value will be either Eq. (6) or Eq. (7):

$$(1 - d) * (d * d) + d * \sqrt{d} \quad (6)$$

or

$$\sqrt{d} \quad (7)$$

4) Once the final output has been evaluated, the process stops.

By using the updated cosine similarity, Enhanced Word2Vec generates word vectors that represent word meanings and their relationships; it is represented as $W2^e$

- GloVe Word Embedding: The GloVe word embedding method [14] is employed to retrieve semantic features from text, P_{Te} . GloVe, or Global Vectors, captures words as a vector space. It is simple to train because of its parallel implementation and assists in capturing word relationships. In the matrix Z , Z_{ij} signifies the count of context word i occurring with word j . The GloVe method is designed to minimize the resulting objective function as defined by Eq. (8).

$$G^{fe} = \text{Minimize} \sum_{i,j=1}^V (W_i^t W_j + B_i + B_j - \log Z_{ij}) \quad (8)$$

The output, G^{fe} from GloVe is a word embedding for each word, W , in the input text P_{Te} .

- Sentence BERT: BERT is a transformer model with an attention mechanism that consumes input text in a cost-effective manner [39]. It is an encoder-only transformer architecture that uses an attention mechanism to process input text and make predictions based on it. BERT processes context in both directions and yields a deeper sense of word meaning. Sentence-BERT creates better sentence-level embeddings, which makes it useful for tasks like retraction classification of published articles by knowing the meaning of entire sentences. BERT's output, Be^{fe} is generally a batch of contextualized word embeddings or sentence embeddings. The overall text-based features are obtained from this feature extraction stage, which employs Enhanced Word2Vec, GloVe, and sentence-BERT to extract features from the punctuated text. The resulting features from this stage are denoted as Fe_{Te} .

$$F_e = [E_{EW2V} || E_{GloVe} || E_{SBERT}] \quad (9)$$

C. Hybrid Classification

After the extraction of text-based features Fe_{Te} , these are used in this classification process. The aim of the classification is to classify retractions. For this, a unique hybrid AAHA-ShLNet strategy is introduced that combines the AAHA-BN-ShNet and LeNet models. The probability outcomes from both classifiers are then combined via soft voting to predict the final classification. Fig. 3 presents a structural depiction of this AAHA-ShLNet strategy.

1) Attention-driven adaptive hybrid activation and batch normalization updated ShuffleNet (AAHA-BN-ShNet): ShuffleNet, a lightweight grouped convolution and channel shuffle operations provide an effective mechanism to learn local feature interactions from dense embedding representation. This approach is being utilized, as the expected inputs are not the raw texts but an embedding matrix. The model aims to achieve maximum efficiency and precision using advanced methods like group convolution (GConv), pointwise group convolution (PWGConv), shuffling, and depthwise convolution (DWConv) [40]. Here, ShuffleNet does not depend on standard "point convolution" in convolution layers but employs GConv, where all the inputs belonging to a group are processed separately. This segregated way ensures that accurate information related to each group is identified, thus enhancing precision. In ShuffleNet, the standard BN normalizes the input layer by layer, which stabilizes the learning process regardless of possible network weight changes. However, it is not necessarily optimized in every layer, and it does not provide smoothness for negative values, which can cause computational inefficiencies or overfitting problems, specifically in article withdrawal reason classification. This overfitting slows down model performance by having the model memorize useless patterns in the data, instead of generalizing well to original inputs.

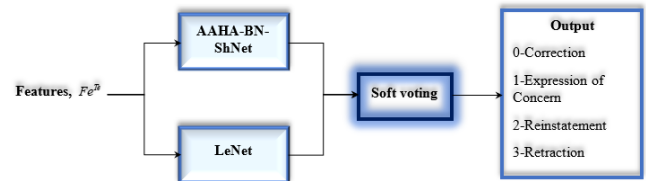


Fig. 3. Structural process of AAHA-ShLNet strategy.

In consideration of such challenges, the ShuffleNet model is improved with the inclusion of an Adaptive Hybrid Activation and Attention-based Batch Normalization (AAHA-BN), which regulates activations in training and enhances the overall performance of the model. Such an enhancement facilitates better generalization of the model through efficient normalization of activations and avoids the susceptibility that comes with traditional BN. The improved ShuffleNet model, tailored specifically for withdrawal text classification in publications, accepts text-based features, Fe_{Te} , as input and goes through a series of layers to efficiently classify. The model starts with a convolution (Conv) layer to generate the initial feature extraction of the input text. This is followed by AAHA-BN normalizing the activations to promote stability, followed by the use of LReLU to impose non-linearity with the inclusion of allowing small negative values. A max pooling layer is then used to decrease dimensionality while retaining important features. The model continues to Stage 2 and Stage 4, which involve Conv layers, AAHA-BN [41], and LReLU, enhancing the feature extraction process in stages. In Modified Stage 3, the model continues to improve feature representation with parallel processing paths involving extra DWConv and AAHA-BN before channel shuffling is used to combine results. After these steps, the features are passed through GAP to convolve all the

feature maps together and are then input into an FC layer to generate the final output. The last output of the AAHA-ShLNet model is indicated as O_{AAHA-BN-ShNet}.

During Modified Stage 3, as indicated in Fig. 4(a) (given in Appendix), the input passes through two parallel paths to boost feature extraction. In the first path, the input initially goes through a Conv layer, followed by the AAHA-BN to normalize activations. The output subsequently passes through an LReLU activation and a DWConv layer, followed by a second AAHA-BN, a Conv layer, and an additional AAHA-BN. Lastly, the output is again activated by LReLU. The second branch uses a less complex flow: it begins with DWConv, then AAHA-BN, Conv, and LReLU activation. Once both branches have been processed, the outputs are concatenated and fed into channel shuffling, which rearranges the features in an efficient manner for the subsequent layer. Fig. 4(b) (given in the Appendix), the input is subjected to channel splitting prior to passing it through a sequence of layers such as a Conv layer, AAHA-BN, LReLU, and DWConv. The output is subsequently processed through AAHA-BN and a Conv layer prior to concatenating with the earlier output. Channel shuffling is the last step in this structure, which efficiently shuffles the features from various channels to enhance feature diversity.

The input is initially divided into several channels for parallel processing to enable the model to learn more varied features for both Stage 2 and Stage 4. The input goes through a Conv layer, followed by BN and a ReLU function. The model then uses a DWConv layer to process the feature representation more precisely, followed by a second BN layer, a Conv layer, and a final BN layer. Then, the result from both the channel split and BN + ReLU branch is concatenated. Then the features undergo channel shuffling to effectively merge the results before proceeding to the next step. The standard BN formula normalizes the activations by adjusting the mean and variance for each feature within a mini-batch, as defined by Eq. (10).

$$\hat{y}_i = \frac{y_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} \quad (10)$$

where, Y_i is the layer feature map, denotes a constant, and implies the mean and variance of y_i . In this model, the standard BN is enhanced, and its updated expression is provided in Eq. (11).

$$\hat{y}_i = \frac{y_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} X P(y) \quad (11)$$

$P(y)$ signifies the Lp pooling with AAHA-BN-based normalization. This pooling method offers better generalization compared to max pooling. Instead of choosing the maximum value, Lp pooling [42] calculates a weighted average of the input within the pooling region [43].

2) *LeNet model*: LeNet is a very effective Deep Neural Network (DNN) model that takes very little computational time. It consists of seven layers, including the output layer [44]. The input layer of the network accepts text-based features FeTe, which are then passed to the next layer after applying a size-normalization operation. The convolutional layer is the next layer, which uses a collection of feature maps or kernels to

select key features of the input. These kernels are collections of square matrices, all with the same weights. The convolutional layer applies learnable kernels over the structured numerical feature representation to identify local feature interactions. In the proposed framework, the input is not an image; rather, the fused textual embeddings are arranged as a fixed-dimensional numerical feature map. The convolution operation therefore enables the LeNet model to extract informative local patterns from the structured embedding representation.

3) *Final classification using soft voting*: Following the withdrawal classification by both AAHA-BN-ShNet and LeNet classifiers [45] on articles, the probabilities from both classifiers' outputs are used as inputs to the soft voting to get the final classification. Soft voting works by evaluating the weighted sum of the prediction probabilities for each class from two classifiers, as expressed in Eq. (12). The class with the highest overall probability is then chosen as the final prediction.

$$\hat{X}_i = \operatorname{argmax}_l \sum_{j=1}^n \omega_j p_{i,l}^j \quad (12)$$

Here, j^{th} classifier weight is implied; the number of classifiers is represented by $n=2$, and is the probability predicted by the j^{th} classifier for the i^{th} sample to be classified as the l^{th} class. By combining the AAHA-BN-ShNet and LeNet classifiers, the proposed hybrid AAHA-ShLNet strategy successfully categorizes withdrawal reasons of articles, which labels the outcome as 0, 1, 2, and 3 (Correction, Expression of Concern, Reinstatement, Retraction).

III. RESULTS

The proposed framework was simulated using PYTHON 3.7. The processor employed was "11th Gen Intel(R) Core (TM) i5-1135G7 @ 2.40GHz 2.42 GHz and the Installed RAM size was 16.0 GB. Moreover, the analysis of withdrawal reason classification on articles was conducted using the Retraction Watch Database from its GitHub repository [46]. This Database aids as a complete repository of retractions—instances where available research articles were withdrawn due to numerous reasons, including "fraud, plagiarism, or ethical concerns." The record contains 52,811 entries, the newest retraction repository. For the experiments, a complete retraction dataset of 52,811 records was utilized, categorized into four diverse classes: Correction (Label 0), Expression of Concern (Label 1), Reinstatement (Label 2), and Retraction (Label 3). The class distribution is highly imbalanced. Retraction constitutes 48,335 records (91.52%), followed by Expression of Concern with 3,092 records (5.85%), Correction with 1,245 records (2.36%), and Reinstatement with 139 records (0.26%). Thus, the minority Reinstatement category contains substantially fewer observations than the majority Retraction category. Because overall accuracy can therefore be dominated by the majority class, macro-averaged F1-score, balanced accuracy, class-wise precision, recall, F1-score, and Matthews correlation coefficient (MCC) are additionally considered in the evaluation.

A. Performance Evaluation

To assess the performance of the framework introduced for the classification of withdrawal articles, a diverse set of performance metrics is used. The selected metrics provide a

comprehensive and balanced evaluation of classification capacity, especially in unbalanced class scenarios. Let T_p denote true positives, T_n be true negatives, F_p false positives, and F_n be false negatives.

1) *Accuracy*: Accuracy measures the general correctness of the classification model as given in Eq. (13).

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \quad (13)$$

Although accuracy provides a general indication of performance, it may be misleading when applied to imbalanced datasets.

2) *Precision*: Precision measures the accuracy of predicted positive instances using Eq. (14).

$$Precision = \frac{T_p}{T_p + F_p} \quad (14)$$

3) *Recall*: Recall, also called sensitivity, is an indicator to accurately measure the effectiveness of the model to identify positive instances, as given in Eq. (15).

$$Recall = \frac{T_p}{T_p + F_n} \quad (15)$$

4) *Specificity*: Specificity measures a model's ability, as per Eq. (16), to correctly identify the negative instances.

$$Specificity = \frac{T_n}{T_n + F_p} \quad (16)$$

The higher value of specificity ensures that fewer incorrect classifications of non-withdrawal cases have been identified.

5) *F1-Score*: The F1-score in Eq. (17) supports a value to be used as a balance between precision and recall:

$$F1 - score = \frac{2 * T_p}{2 * T_p + F_p + F_n} \quad (17)$$

It is a crucial value for unbalanced datasets.

6) *Matthews Correlation Coefficient (MCC)*: As per the MCC value, Eq. (18) provides a balanced measure for considering all elements of the confusion matrix:

$$MCC = \frac{T_p * T_n - F_p * F_n}{\sqrt{(T_p + F_p)(T_p + F_n)(T_n + F_p)(T_n + F_n)}} \quad (18)$$

The MCC value ranges from -1 (total disagreement) to +1 (perfect prediction).

7) *Negative Predicted Value (NPV)*: The NPV in Eq. (19) helps to measure the proportion of correctly predicted negative instances:

$$NPV = \frac{T_n}{T_n + F_n} \quad (19)$$

8) *False Positive Rate (FPR)*: The FPR value given in Eq. (20) helps quantify the proportion of negative instances incorrectly classified as positive:

$$FPR = \frac{F_p}{F_p + T_n} \quad (20)$$

Lower FPR values indicate reduced false alarms.

9) *False Negative Rate (FNR)*: The FNR value shows the exact proportion of all positive instances incorrectly classified as negative instances, calculated using Eq. (21):

$$FNR = \frac{F_n}{F_n + T_p} \quad (21)$$

Minimizing FNR is critical to avoid missing relevant withdrawal cases.

For rigorous validation, the proposed framework is evaluated using multiple training ratios and compared with baseline models. In addition to this, statistical measures including mean, median, standard deviation, and range, together with ablation analysis, are used to assess the robustness and reliability of the framework.

B. Comparative Performance Analysis

To evaluate the efficacy of the proposed model, a comparative analysis was conducted against several established models, including CNN, Bi-LSTM, GoogLeNet, Bi-GRU, LSTM, RNN, and AlexNet. The results showed that the model achieved superior performance in four-class post-publication notice classification, achieving an accuracy of 94.3%. In contrast, baseline models exhibit comparatively lower accuracy levels. Furthermore, in 90% of the training data, AAHA-ShLNet achieves a precision score of 0.885, highlighting its effectiveness. Baseline models report lower precision values, including CNN (0.836), Bi-LSTM (0.819), GoogLeNet (0.848), Bi-GRU (0.839), LSTM (0.859), RNN (0.826), and AlexNet (0.829).

The Negative Predictive Value (NPV) analysis demonstrates the robustness of the proposed model in identifying and classifying the removed articles. The model achieved the highest NPV of 0.927 in 60% of the training data, indicating a strong predictive capacity. The model was further improved from 0.936 to 0.943 with an increase in data from 70% to 90%, respectively. Hence, these results confirm the consistent enhancement of classification reliability with an increase in the dataset.

The effective withdrawal of categorization requires minimal negative performance indicators, such as false positive rates (FPR). The proposed model achieved the lowest FPR value of 0.032 in the training data of 90%, demonstrating a significant reduction in errors. In contrast, conventional models show higher FPR scores with Bi-LSTM (0.063), RNN (0.061), AlexNet (0.060), CNN (0.057), and Bi-GRU (0.059). The results highlight the proposed model's competence in minimizing classification errors. In addition, the Word2Vec component integrated an adaptive scaling factor and log-exp transformation to regulate similarity measures according to the variability of the dataset, thus producing a more balanced similarity measure less sensitive to outliers. The detailed description of all performance metrics compared to the traditional system is given in Table II. The statistical analysis is expanded to include macro-averaged precision, recall, and F1-score, balanced accuracy, class-wise precision/recall/F1, and a four-class confusion matrix. The performance of the minority Reinstatement class is explicitly reported rather than being represented only through aggregate accuracy.

TABLE II. COMPARATIVE PERFORMANCE ANALYSIS OF PROPOSED AND BASELINE MODELS

Model	Accuracy	Precision	FPR	FNR
AAHA - ShLNet	0.943	0.885	0.032	0.157
CNN	0.920	0.836	0.057	0.180
Bi-LSTM	0.913	0.819	0.063	0.195
GoogLeNet	0.920	0.848	NR	0.180
Bi-GRU	0.915	0.839	0.059	0.190
LSTM	0.919	0.859	NR	0.175
RNN	0.915	0.826	0.061	0.190
AlexNet	0.915	0.829	0.060	NR

C. Statistical Evaluation

To assess the effectiveness of the framework, a comprehensive quantitative evaluation was conducted using statistical measures of accuracy. The results, summarized in Table III, provide a comparative analysis of the introduced model compared to several advanced deep learning models and the baseline, such as CNN, LSTM, Bi-LSTM, GoogLeNet, Bi-GRU, RNN, and AlexNet. Statistical evaluation includes minimum, mean, standard deviation, median, and maximum accuracy values. Thus, a comprehensive understanding of the performance of the model under different experimental conditions is possible. Among all evaluated models, the proposed model demonstrates superior and consistent performance. More specifically, the mean accuracy for the proposed architecture was estimated at 0.943, which is higher than the mean accuracy of all baseline models, ranging from 0.913 to 0.920. It should be noted that the model demonstrates an ability to achieve fairly stable prediction performance across multiple iterations and datasets. In addition, the maximum accuracy of 0.972 provided by the proposed model outperforms all other baselines, including LSTM (0.950), GoogLeNet (0.946), and CNN (0.944). Thus, the suggested architecture allows for the highest classification performance under optimal conditions. As for the model's robustness, its relatively small standard deviation equals 0.019, which signifies that the performance remains quite stable and there are no major differences between different experiments. The median accuracy of 0.941 supports this statement, as it almost coincides with the mean value. The baseline models not only have a lower average accuracy than the proposed model, but also show a wider range of performance, as shown by their higher or similar standard deviation values. This means that traditional architectures may have trouble generalizing well when there are complex and unbalanced withdrawal data. Combining multi-level feature representations makes the model better at capturing both semantic and contextual information. In addition, using Adaptive Hybrid Activation with Attention-based Batch Normalization (AAHA-BN) improves feature normalization and makes training more stable. The hybrid classification strategy that combines AAHA-BN-ShNet and LeNet also helps the decision to be more general and stronger. In general, statistical analysis confirms that the proposed framework achieves higher precision, improved stability, and superior generalization capability compared to existing approaches, making it highly suitable for reliable four-class post-publication notice classification in the scientific literature.

TABLE III. STATISTICAL EVALUATION OF ACCURACY

Model	Min	Mean	SD	Median	Max
AAHA-ShLNet	0.919	0.943	0.019	0.941	0.972
CNN	0.891	0.920	0.018	0.920	0.944
LSTM	0.891	0.919	0.022	0.918	0.950
Bi-LSTM	0.884	0.913	0.020	0.916	0.938
GoogLeNet	0.896	0.920	0.019	0.919	0.946
Bi-GRU	0.887	0.915	0.021	0.916	0.943
RNN	0.889	0.915	0.019	0.916	0.939
AlexNet	0.888	0.915	0.019	0.916	0.940
TF-IDF + Linear SVM	0.753	0.86	0.052	0.855	0.88

To examine the effectiveness of the proposed model, a comparison is made with two distinct variations, one with enhanced ShuffleNet and the other without enhanced Word2vec. The details of these comparisons are summarized in Table IV. The results of the ablation evaluation clearly highlight the superior performance of the AAHA-ShLNet model. Specifically, the scheme achieved the maximum sensitivity rate of 0.843, outperforming AAHA-ShLNet without enhanced ShuffleNet (0.747) and AAHA-ShLNet without enhanced Word2vec (0.699). In addition, the AAHA-ShLNet scheme acquired the lowest FPR score of 0.032, indicating reduced error values. In contrast, the proposed model with existing ShuffleNet and AAHA-ShLNet with existing Word2vec recorded relatively higher FPR values of 0.084 and 0.100, respectively.

TABLE IV. ABLATION STUDY OF THE PROPOSED FRAMEWORK

Measure	AAHA-ShLNet	w/o Enhanced ShuffleNet	w/o Enhanced Word2Vec
Accuracy	0.943	0.916	0.900
MCC	0.791	0.663	0.599
Sensitivity	0.843	0.747	0.699
FNR	0.157	0.253	0.301
Specificity	0.922	0.874	0.850
NPV	0.948	0.916	0.900
Precision	0.885	0.748	0.700
F-Measure	0.843	0.747	0.700
FPR	0.032	0.084	0.100

IV. DISCUSSION

The proposed experimental work provides a comprehensive assessment of four-class post-publication notice classification using multiple training-data configurations and class-sensitive performance measures, as compared to the traditional approaches on multiple evaluation metrics, thus validating its effectiveness for an automated classification of article withdrawal reasons. An accuracy of 94.3% has been achieved, with an improvement across all the key metrics such as precision, recall, specificity, and predictive reliability. The hybrid structure enables the model to effectively manage the complex and imbalanced retraction datasets. This improved performance is likely due to the fusion of multilevel semantic

feature extraction and hybrid classification mechanisms. The withdrawal notices often have inconsistent reporting structures, leading to an ambiguous understanding along with domain-specific terminologies, making the automated interpretation highly challenging. Similar concerns have been highlighted by Rao et al. [22], Hwang [12], and Vuong [30]. The result confirms that the hybrid approach has a better generalization capability than standalone models for handling the heterogeneous textual data.

The convolutional components in the proposed framework do not operate on image pixels or visual content. Instead, they operate on a structured numerical feature map constructed from the fused textual representations. Enhanced Word2Vec, GloVe, and Sentence-BERT generate complementary numerical representations at different levels of linguistic abstraction. These representations are concatenated to form a unified feature vector and subsequently arranged into a fixed-dimensional numerical feature map. The two-dimensional arrangement is therefore a computational representation rather than an assumption that the input possesses visual semantics. Convolutional filters are used to learn local interactions among neighboring embedding dimensions and feature channels. Grouped convolution and channel-shuffling operations in ShuffleNet provide parameter-efficient learning of these interactions, while LeNet provides a complementary convolutional pathway. Thus, the use of convolution is motivated by structured feature interaction learning rather than by treating textual data as an image.

The semantic understanding as well as the contextual representation has improved by integrating advanced embedding techniques like Enhanced Word2Vec, GloVe, and Sentence-BERT. On one side, the traditional approaches often fail to capture long-range contextual dependencies and their semantic relationships, while on the other hand, the proposed multi-representation embedding strategy enabled the model to capture the linguistic patterns at the word, sentence, and even contextual level simultaneously. The resulting tensor is a structured feature representation rather than a natural image, and the convolutional layers are employed as computational feature extractors over locally arranged embedding dimensions. The overall robustness of the model against noisy and unstructured text has been improved by utilizing modified cosine similarity with enhanced word2Vec by reducing the sensitivity towards outliers and variable data distribution. A similar observation has been demonstrated by Yusof et al. [37], Shepperd & Yousefi [34], and Rao et al. [22]. Therefore, this fusion has substantially improved the classification reliability for complex scientific datasets comprising ambiguous textual patterns.

The incorporation of Adaptive Hybrid Activation with Attention-based Batch Normalization (AAHA-BN) and the hybrid soft voting classification strategy enhanced the feature normalization and stabilized the learning process of the model by effectively regulating the activation distribution during the training process, which is another important contribution of this study. This combination supports the reduction of overfitting issues and improvement in convergence stability while processing imbalanced datasets. For integrating hierarchical feature extraction adaptability with computational efficiency, the hybrid combination of AAHA-BN-ShNet and LeNet was found to be effective. The predictive robustness has been

enhanced by the aggregation of complementary classifier outputs into a unified decision-making framework using a soft voting mechanism. Further, the ablation study performed has clearly demonstrated a noticeable degradation in classification metrics such as accuracy, sensitivity, and specificity in the case of removing any of the above mechanisms from the model. A similar discussion has been presented by Baker [28] that shows the significance of normalization and robust features in deep learning approaches defined by Sharma et al. [36] and Chan et al. [32]. Hence, this collective approach is best suited for reliable classification tasks.

The proposed framework thus offers a scalable and efficient solution for automatic scientific integrity monitoring and identification of unreliable publications across multiple research domains. The recent gradual hike in article withdrawals globally highlights the urgent requirement for such intelligent systems that are capable of supporting publishers, indexing agencies, policymakers, and researchers in identifying the root cause of withdrawal more effectively. This systematic approach helps reduce the propagation of misinformation and further improves transparency in scientific communications. Haunschild & Bornmann [23] and Pastor-Ramón et al. [24] have highlighted similar concerns in their studies and emphasized the requirement of such a cohesive and robust system. With these promising results, there are certain limitations with this work. First, the study heavily relies on the Retraction Watch Database, and thus the completeness, terminologies, and reporting practices of post-publication notices may vary from publishers, journals, as well as disciplines. Consequently, bias may occur depending on the data source. Second, the existing data imbalance estimation of minority-class performance remains more difficult when class support is limited. Third, temporal or domain shifts can impact the overall result and may therefore require additional validation on future records and datasets from independent sources. Fourth, a requirement for some leakage control mechanism to avoid similar records in analysis. Future work may include utilizing explainable AI techniques for supporting multilingual withdrawal notices along with multi-label classification to further improve the scalability and real-world applicability for automatic article withdrawal analysis. External validation across independent databases and multilingual or cross-publisher collections is therefore an important direction for future work.

V. CONCLUSION

The work presents a novel hybrid framework, termed AAHA-ShLNet, for the four-class classification of post-publication editorial notices. The proposed approach integrates a structured pipeline consisting of text preprocessing, multi-level feature extraction, and hybrid classification. First, data preprocessing is performed to remove redundant punctuation while keeping the essential linguistic structure. In addition, semantic and contextual representations are captured using enhanced embedding techniques such as Enhanced Word2Vec, GloVe, and Sentence-BERT, allowing exhaustive feature learning from collated textual data. For classification, a hybrid strategy is employed that combines AAHA-BN-ShNet and LeNet, where the outputs of both models are aggregated using a soft voting mechanism to improve predictive robustness. The experimental results show that the proposed framework is competitive with conventional models. In particular, the model

achieved a precision of 0.885, better than the traditional baseline approaches such as CNN(0.836), Bi-LSTM(0.819), GoogLeNet(0.848), Bi-GRU(0.839), LSTM(0.859), RNN(0.826), and AlexNet (0.829). Performance gains are attributed to the integration of complementary feature representations and the enhanced normalization strategy incorporated within the AAHA-ShLNet architecture. Furthermore, the hybrid decision fusion mechanism contributes to improved generalization and reduced classification errors. In general, the proposed framework provides an effective and scalable solution to identify withdrawal-related patterns. By allowing for more accurate detection of unreliable publications, the model supports informed decision-making, reduces the risk of spreading misinformation, and contributes to maintaining the integrity and reliability of scientific communication. In future work, the framework will be extended to support multi-label and fine-grained classification scenarios where multiple underlying causes can be associated with a single article. Furthermore, explainable artificial intelligence (XAI) techniques can be integrated to improve the interpretations and transparency of improved classification decisions. The framework demonstrates strong potential for real-world deployment in large-scale scientific monitoring systems.

ACKNOWLEDGMENT

The authors would like to acknowledge their university for providing the infrastructure required to carry out the experiments.

DECLARATION OF GENERATIVE AI USAGE

During the preparation of this manuscript, no generative artificial intelligence tools were used.

REFERENCES

- [1] B. A. Nosek, D. B. Allison, K. H. Jamieson, M. McNutt, A. B. Nielsen, and S. M. Wolf, "A framework for assessing the trustworthiness of scientific research findings," *Proceedings of the National Academy of Sciences*, vol. 123, no. 6, Art. no. e2536736123, 2026.
- [2] Z. Syed, F. Syed, L. Thabane, and M. Rodrigues, "COVID-19 retracted publications on Retraction Watch: A systematic survey of their preprints and citations," *Heliyon*, vol. 9, no. 4, 2023.
- [3] I. Campos-Varela and A. Ruano-Raviña, "Misconduct as the main cause for retraction: A descriptive study of retracted publications and their authors," *Gaceta Sanitaria*, vol. 33, pp. 356–360, 2019.
- [4] X. Zheng, J. Chen, A. Tollas, and C. Ni, "The effectiveness of peer review in identifying issues leading to retractions," *Journal of Informetrics*, vol. 17, no. 3, Art. no. 101423, 2023.
- [5] Q. H. Vuong, V. P. La, M. T. Ho, T. T. Vuong, and M. T. Ho, "Characteristics of retracted articles based on retraction data from online sources through February 2019," *Science Editing*, vol. 7, no. 1, pp. 34–44, 2020.
- [6] C. Chen, Z. Hu, J. Milbank, and T. Schultz, "A visual analytic study of retracted articles in scientific literature," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 2, pp. 234–253, 2013.
- [7] X. Liu, C. Wang, D. Z. Chen, and M. H. Huang, "Exploring perception of retraction based on mentioned status in post-retraction citations," *Journal of Informetrics*, vol. 16, no. 3, Art. no. 101304, 2022.
- [8] F. C. Fang, R. G. Steen, and A. Casadevall, "Misconduct accounts for the majority of retracted scientific publications," *Proceedings of the National Academy of Sciences*, vol. 109, no. 42, pp. 17028–17033, 2012.
- [9] C. Candal-Pedreira, C. Guerra-Tort, A. Ruano-Ravina, F. Freijedo-Fariñas, J. Rey-Brandariz, J. S. Ross, and M. Pérez-Ríos, "Retracted papers originating from paper mills: A cross-sectional analysis of references and citations," *Journal of Clinical Epidemiology*, vol. 172, Art. no. 111397, 2024.
- [10] C. Brook, N. Lynøe, A. Eriksson, and D. Balding, "Retraction of a peer-reviewed article suggests ongoing problems with Australian forensic science," *Forensic Science International: Synergy*, vol. 3, Art. no. 100208, 2021.
- [11] E. Sung, J. L. Ramirez, A. Z. Becerra, E. Jaramillo, W. J. Gasper, P. Schneider, M. S. Conte, and J. C. Iannuzzi, "Identifying disruptive publications and changing paradigms in vascular surgery," *JVS-Vascular Insights*, vol. 2, Art. no. 100036, 2024.
- [12] S. Y. Hwang, D. K. Yon, S. W. Lee, M. S. Kim, J. Y. Kim, L. Smith, A. Koyanagi, M. Solmi, A. F. Carvalho, E. Kim, and J. I. Shin, "Causes for retraction in the biomedical literature: A systematic review of studies of retraction notices," *Journal of Korean Medical Science*, vol. 38, no. 41, Art. no. e333, 2023.
- [13] P. Sebo and M. Sebo, "Comparing the performance of Retraction Watch Database, PubMed, and Web of Science in identifying retracted publications in medicine," *Accountability in Research*, vol. 33, no. 2, 2026.
- [14] B. F. Kocyigit, A. Zhaksylyk, A. Akyol, and M. Yessirkepov, "Characteristics of retracted publications from Kazakhstan: An analysis using the Retraction Watch Database," *Journal of Korean Medical Science*, vol. 38, no. 46, Art. no. e390, 2023.
- [15] P. Khurana, K. Sharma, and Z. Uddin, "Unraveling retraction dynamics in COVID-19 research: Patterns, reasons, and implications," *Accountability in Research*, vol. 32, no. 7, pp. 1241–1264, 2025.
- [16] M. P. Contreras-Muñoz, C. Zahn-Muñoz, E. Solís-Albanese, and E. Martínez-Rojas, "Analysis of retractions in nursing from publications between 2000 and 2024: A bibliometric analysis using Retraction Watch," *Nursing Reports*, vol. 15, no. 10, Art. no. 349, 2025.
- [17] X. Rossello, S. Raposeiras-Roubín, R. Latini, A. Domínguez-Rodríguez, J. A. Barrabés, P. L. Sánchez, M. Anguita, F. Fernández-Vázquez, D. Pascual-Figal, J. M. de la Torre Hernández, and S. Ferraro, "Rationale and design of the pragmatic clinical trial tREatment with Beta-blockers after myOcardial infarction withOut reduced ejection fracTion (REBOOT)," *European Heart Journal-Cardiovascular Pharmacotherapy*, vol. 8, no. 3, pp. 291–301, 2022.
- [18] R. A. Okay, B. F. Kocyigit, A. B. Kumar, M. Yessirkepov, and H. E. Sumbul, "Fifty years of retracted medical publications from 1975 to 2024: A comprehensive analysis of trends, reasons, and countries using the Retraction Watch Database," *Journal of Korean Medical Science*, vol. 40, no. 46, Art. no. e300, 2025.
- [19] J. I. de Granda-Orive, C. A. Jiménez-Ruiz, R. Aleixandre-Benavent, D. López-Padilla, C. Rábade-Castedo, M. J. Gómez, and A. Alonso-Arroyo, "Trends and characteristics of retracted articles in the smoking field: An observational study," *Open Respiratory Archives*, Art. no. 100496, 2025.
- [20] P. Sharma, B. Sharma, A. Reza, K. K. Inampudi, and R. K. Dhamija, "A systematic review of retractions in biomedical research publications: Reasons for retractions and their citations in Indian affiliations," *Humanities and Social Sciences Communications*, vol. 10, no. 1, Art. no. 597, 2023.
- [21] V. Bhat, S. Kaushik, A. Kumar, S. Bellur, and A. Kalra, "Retractions and withdrawals in clinical cardiovascular literature: Analysis of the Retraction Watch Database," *JACC: Advances*, vol. 4, no. 8, Art. no. 101802, 2025.
- [22] D. Rao, J. Young, T. Dietterich, and C. Callison-Burch, "WithdrarXiv: A large-scale dataset for retraction study," 2024.
- [23] R. Haunschild and L. Bornmann, "Can tweets be used to detect problems early with scientific papers? A case study of three retracted COVID-19/SARS-CoV-2 papers," *Scientometrics*, vol. 126, no. 6, pp. 5181–5199, 2021.
- [24] E. Pastor-Ramón, I. Herrera-Peco, O. Agirre, M. García-Puente, and J. M. Morán, "Improving the reliability of literature reviews: Detection of retracted articles through academic search engines," *European Journal of Investigation in Health, Psychology and Education*, vol. 12, no. 5, pp. 458–464, 2022.

- [25] T. Zilberman, I. Margalit, D. Yahav, and N. Tau, "Retracted publications in infectious diseases and clinical microbiology literature: An analysis using the Retraction Watch Database," *Clinical Microbiology and Infection*, vol. 29, no. 11, pp. 1454.e1–1454.e8, 2023.
- [26] C. J. Bakker, E. E. Reardon, S. J. Brown, N. Theis-Mahon, S. Schroter, L. Bouter, and M. P. Zeegers, "Identification of retracted publications and completeness of retraction notices in public health," *Journal of Clinical Epidemiology*, vol. 173, Art. no. 111427, 2024.
- [27] Q. Qi, J. Huang, Y. Wu, Y. Pan, J. Zhuang, and X. Yang, "Recent trends: Retractions of articles in the oncology field," *Heliyon*, vol. 10, no. 12, Art. no. e33007, 2024.
- [28] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature*, vol. 533, no. 7604, pp. 452–454, 2016.
- [29] J. P. A. Ioannidis, J. Baas, R. Klavans, and K. W. Boyack, "A standardized citation metrics author database annotated for scientific field," *PLOS Biology*, vol. 17, no. 8, Art. no. e3000384, 2019.
- [30] Q. H. Vuong, "Reform retractions to make them more transparent," *Nature*, vol. 582, no. 7811, p. 149, 2020.
- [31] N. S. L. Yeo-Teh and B. L. Tang, "An alarming retraction rate for scientific publications on Coronavirus Disease 2019 (COVID-19)," *Accountability in Research*, vol. 28, no. 1, pp. 47–53, 2021.
- [32] A. W. Chan, F. Song, A. Vickers, T. Jefferson, K. Dickersin, P. C. Göttsche, H. M. Krumholz, D. Ghersi, and H. B. van der Worp, "Increasing value and reducing waste: Addressing inaccessible research," *The Lancet*, vol. 383, no. 9913, pp. 257–266, 2014.
- [33] N. Mascato Fontañá, C. Candal-Pedreira, G. García, J. S. Ross, A. Ruano-Ravina, and L. Martin-Gisbert, "Identifying common patterns in journals that retracted papers from paper mills: A cross-sectional study," *Research Integrity and Peer Review*, vol. 10, no. 1, Art. no. 21, 2025.
- [34] M. Shepperd and L. Yousefi, "An analysis of retracted papers in computer science," *PLOS ONE*, vol. 18, no. 5, Art. no. e0285383, 2023.
- [35] J. P. Tennant, F. Waldner, D. C. Jacques, P. Masuzzo, L. B. Collister, and C. H. Hartgerink, "The academic, economic and societal impacts of open access: An evidence-based review," *F1000Research*, vol. 5, Art. no. 632, 2016.
- [36] K. Sharma and P. Khurana, "Retracted citations and self-citations in retracted publications: A comparative study of plagiarism and fake peer review," 2025.
- [37] N. H. Mohd Yusof, N. A. Mohd Subha, N. Hamzah, S. Buyamin, and M. A. Shamsudin, "Logical entity weighting and modified cosine similarity for automated triples extraction via HMM-BERT in knowledge-based maintenance systems for the cement industry," 2025.
- [38] L. Sahu and B. R. Mohan, "An improved K-means algorithm using modified cosine distance measure for document clustering using Mahout with Hadoop," in *Proc. International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, 2014, pp. 1–5.
- [39] A. A. Jamjoom, H. Karamti, M. Umer, S. Alsubai, T. H. Kim, and I. Ashraf, "RoBERTaNet: Enhanced RoBERTa transformer-based model for cyberbullying detection with GloVe features," *IEEE Access*, vol. 12, pp. 58950–58959, 2024.
- [40] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6848–6856.
- [41] A. A. Ziaee and E. Çano, "Batch Layer Normalization: A New Normalization Layer for CNNs and RNNs," in *Proceedings of the 6th International Conference on Advances in Artificial Intelligence (ICAAI '22)*, Birmingham, U.K., Oct. 2022, pp. 40–49. doi: 10.1145/3571560.3571566.
- [42] H. Gholamalinezhad and H. Khosravi, "Pooling Methods in Deep Neural Networks: A Review," *CoRR*, vol. abs/2009.07485, 2020.
- [43] C. C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of Trends in Practice and Research for Deep Learning," *CoRR*, vol. abs/1811.03378, 2018.
- [44] H. Gholamalinezhad, H. Khosravi, "An improved LeNet deep neural network model for Alzheimer's disease classification using brain magnetic resonance images," *IEEE Access*, vol. 9, pp. 161194–161207, 2021.
- [45] K. Akyol, E. Uçar, Ü. Atila, and M. Uçar, "An ensemble approach for classification of tympanic membrane conditions using soft voting classifier," *Multimedia Tools and Applications*, vol. 83, no. 32, pp. 77809–77830, 2024.
- [46] *Retraction Watch Database*. Retraction Watch, Center for Scientific Integrity, 2026. Available: <https://retractionwatch.com/retracted-paper-database/>.

APPENDIX

TABLE V. COMPARATIVE ANALYSIS OF RECENT STUDIES

Scope	Methodology	Findings	Identified Gaps
Life Science Journals [8]	Retrospective Analysis	delay between publication and withdrawal	Delayed detection mechanisms
Engineering and Applied Sciences[26]	Case Study	Reproducibility	Lack of reproducibility focused categories
Global Biomedical Literature[27]	Bibliometric Analysis	Misconduct-driven withdrawals in developing regions	Absence of a specific taxonomy
Retraction Watch[28]	Statistical Analysis	Data related issues dominate with rise in multi-cause cases	Lack of standardized taxonomy
COVID-19 Publications[29]	Comparative Analysis	Increased withdrawal due to rapid publishing	Temporal Bias
Multicenter Clinical Trials[30]	Meta Analysis	Protocol deviations and reporting inconsistencies	Insufficient process oriented classification
Nursing & Medical Education[14]	Systematic Review	Ethical violations reported 80% cases	Domain specific bias limits generalization
Multidisciplinary Datasets[31]	NLP-based machine learning	Multi label classification improves scalability	Dependent on notice quality
Cross-disciplinary Journals [32]	Mixed method Analysis	Distinction between error vs misconduct	Inconsistent reporting formats
Open-Access Journals[33]	Cross-Sectional Study	Higher visibility linked to Faster detection	Bias towards indexed/open access sources
AI- assisted scholarly datasets[34]	Classification	Improved Accuracy in detecting nuanced reasons	Limited model explainability

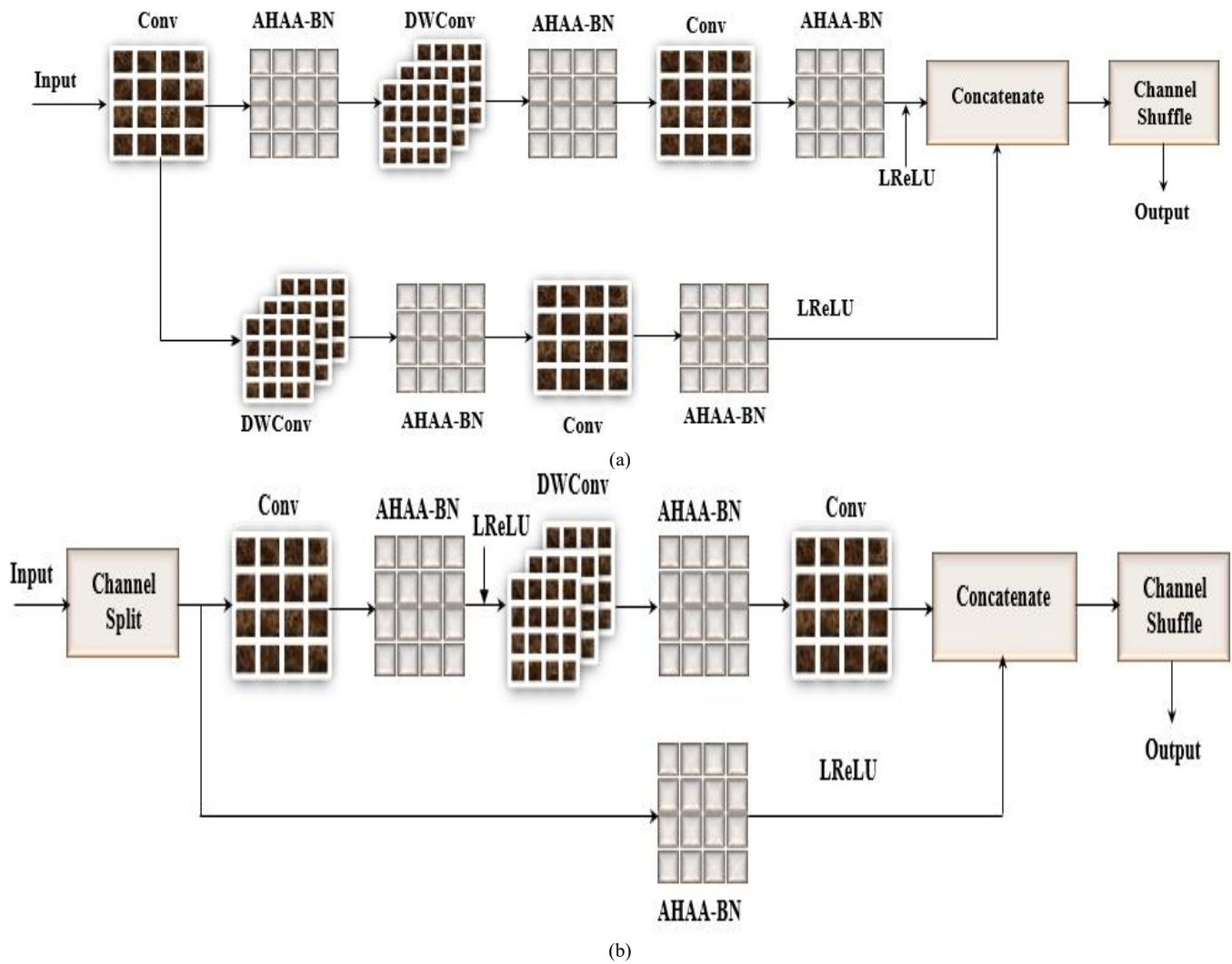


Fig. 4. Depiction of the modified Stage 3 in the AAHA-BN-ShNet model: (a) Non-layered structure; (b) Layered structure.