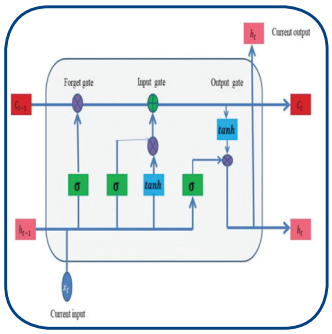
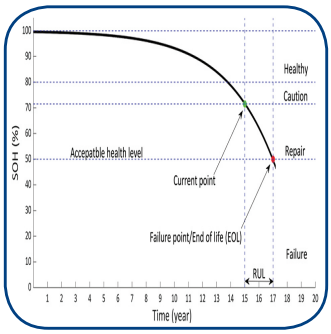
Volume 17, Number 1, 2026 ISSN 1847-6996

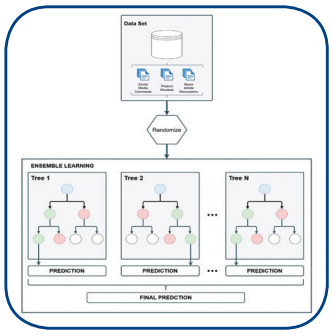
FERIT
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY **OSIJEK**

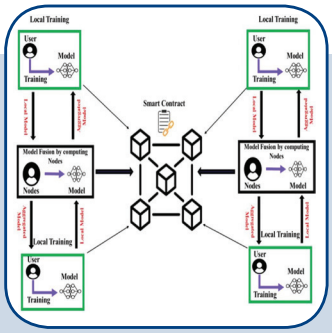
IJECES International Journal of Electrical and Computer Engineering Systems

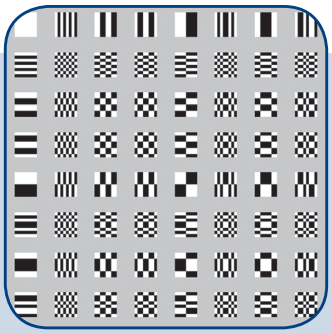
International Journal of Electrical and Computer Engineering Systems

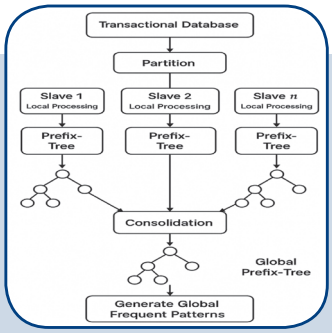












ISSN: 1847-6996

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek, Croatia

Osijek, Croatia | Volume 17, Number 1, 2026 | Pages 1 - 81

The International Journal of Electrical and Computer Engineering Systems is published with the financial support of the Ministry of Science and Education of the Republic of Croatia

CONTACT
International Journal of Electrical and Computer Engineering Systems (IJECES)

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b, 31000 Osijek, Croatia
Phone: +38531224600, Fax: +38531224605
e-mail: ijeces@ferit.hr

Subscription Information
The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.
Giro account: 2390001 - 1100016777,
Croatian Postal Bank

Journal is referred in:

- Scopus
- Web of Science Core Collection (Emerging Sources Citation Index - ESCI)
- Google Scholar
- CiteFactor
- Genamics
- Hrčak
- Ulrichweb
- Reaxys
- Embase
- Engineering Village

Bibliographic Information
Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003
Published: quarterly
Circulation: 300

IJECES online
https://ijeces.ferit.hr

TABLE OF CONTENTS

Leveraging Word2Vec-Enhanced CNN-LSTM Hybrid Architecture for Sentiment Analysis in E-Commerce Product Reviews

Original Scientific Paper

Kosala Natarajan*
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology,
Jeppiar Nagar, Chennai, Tamil Nadu - 600119, India
kosala.nataraj@gmail.com

Nirmalrani V
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology,
Jeppiar Nagar, Chennai, Tamil Nadu - 600119, India
nirmalrani.it@sathyabama.ac.in

Gowri S
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology,
Jeppiar Nagar, Chennai, Tamil Nadu - 600119, India.
gowri.it@gmail.com

Ramya G Franklin
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology,
Jeppiar Nagar, Chennai, Tamil Nadu - 600119, India.
mikella.prabu@gmail.com

Poornima D
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology,
Jeppiar Nagar, Chennai, Tamil Nadu - 600119, India.
poorniramesh2011@gmail.com

Jabez J
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology,
Jeppiar Nagar, Chennai, Tamil Nadu - 600119, India.
jabezme@gmail.com

*Corresponding author

Abstract *– The amalgamation of machine learning (ML) techniques and natural language processing (NLP) is leveraged to evaluate the sentiment of textual input. With the increasing popularity of e-commerce platforms like Amazon, product reviews have emerged as an essential source of information for potential purchasers, providing insights into product quality and performance from the consumers' viewpoints. This study aims to systematically organize and analyze customer opinions to effectively capture consumer sentiment based on product reviews. In this study, we propose a deep learning framework that combines a stacked 1D convolutional layer (CNN) with a Long Short-Term Memory (LSTM) network, using pre-trained Word2Vec embedding as fixed input representations. Evaluated on a large Amazon product review dataset, our model — StackedCNN-LSTM-W2V — achieves a classification accuracy of* **99 %**, *outperforming traditional CNN, LSTM, and logistic regression baselines.*

Keywords*: Sentiment analysis, Amazon product reviews, StackedCNN-LSTM, Text classification, Deep learning, Word embedding*

Received: April 18, 2025; Received in revised form: August 13, 2025; Accepted: August 20, 2025

1. INTRODUCTION

Sentiment analysis (SA) is a branch of Natural Language Processing (NLP) that utilizes machine learning methods to evaluate textual information. It has garnered considerable interest from researchers and developers owing to its efficacy in assessing the polarity of textual content—positive, negative, or neutral. SA has been extensively utilised across many text data types, including product reviews on e-commerce platforms such as Amazon.

Amazon has become a central hub for product feedback, where consumers share their experiences with various items, from electronics to household goods. These reviews serve as valuable insights for both potential buyers and businesses. Customers rely on product reviews to make informed purchasing decisions, while companies analyze sentiment trends to improve product quality, enhance customer satisfaction, and refine marketing strategies [1].

With the transition from traditional to digital marketing, consumer behavior has changed significantly, but word of mouth remains crucial. Platforms like Amazon, Meesho, and Ajio enable customers to share their insights on products and services, influencing other buyers and business strategies. Likewise, major social platforms contribute significantly to shaping consumer perception and corporate reputations.

Customers can acquire significant information regarding a product's quality by perusing the reviews, hence facilitating more informed purchasing judgments. Corporations can utilise this feedback to gauge client happiness and enhance their products or services [2]. Large volumes of customer reviews make manual analysis impractical and inefficient.

The vast amount of unstructured textual data requires transformation into a computable format for efficient processing. Sentiment analysis provides a feasible solution by utilizing text mining techniques to extract subjective information and classify sentiment polarity [3]. Traditional approaches struggle to identify intricate patterns and contextual nuances in reviews, making deep learning(DL) techniques a more viable approach.

Numerous approaches exist for processing product reviews, with DL being one of the most used methods, employing neural networks with numerous layers. DL has been widely utilised in numerous academic domains, including NLP and SA.

Deep learning is chosen for analyzing Amazon reviews due to its ability to capture complex text patterns, improving sentiment classification accuracy. In contrast to conventional machine learning methods, deep learning models exhibit enhanced efficacy in comprehending context, managing long-range dependencies, and deriving significant insights from user feedback.

Deep learning approaches, particularly LSTM and CNN, have shown remarkable improvements in sentiment classification [4]. Traditional machine learning models, though effective, are limited in capturing long-term dependencies and complex linguistic structures. Deep learning models, particularly those combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, have demonstrated strong performance in extracting both spatial and temporal features. This study proposes an advanced model integrating deep stacked CNN layers, frozen Word2Vec embedding, and a robust regularization strategy, specifically optimized for Amazon product reviews, a domain known for its verbose, informal, and sentimentally complex text.

The key contributions of this research are as follows:

- **Utilization of Word2Vec for Feature Extraction**: This study uses Word2Vec to transform text into numerical vectors, capturing rich semantic meanings from review data
- **Development of a Hybrid CNN-LSTM Model**: The proposed StackedCNN-LSTM -WordtoVec model integrates CNN and LSTM layers to extract spatial and sequential features from review text.
- **Feature Extraction using CNN**: CNN is utilized to identify important textual patterns, such as word combinations and sentiment cues, contributing to enhanced feature learning.
- **Context Understanding with LSTM**: By integrating LSTM, the model can learn and retain extended contextual and sequential patterns.

2. RELATED WORKS

Sharma *et. al.* [5] demonstrated that Word2Vec embedding, particularly when kept frozen, outperforms Glove and FastText in deep learning models such as CNN-LSTM when applied to product review datasets. This was attributed to Word2Vec's ability to retain meaningful local word relationships, which is critical for sentiment classification.

Zarei et al. [6] showed that Word2Vec outperforms other embeddings in tasks involving longer English texts, such as e-commerce or social media reviews. Their study highlighted that while glove captures global word co-occurrence, Word2Vec is more effective at encoding local semantic structure, which helps in capturing subtle sentiment nuances.

Hashmi et al. [7] share a similar goal with our research, enhancing sentiment classification performance on Amazon product reviews using various embedding strategies and classification algorithms. While their work adopts a hybrid modelling framework integrating multiple machine learning and deep learning methods, our study emphasizes performance analysis using carefully selected embedding techniques and overfitting control strategies. Both approaches aim to improve classification accuracy and interpretability on real-world e-commerce review data. BERT model reached an accuracy of 89%. The comparative results help validate the effectiveness of different modeling strategies, offering valuable insights into the advantages and drawbacks of different sentiment analysis techniques.

Shamal et al. [8] employed the LSTM model in their research, yielding enhanced outcomes for SA tasks. Furthermore, Guner et al. [9] established that the LSTM outshone competing models in terms of accuracy for binary SA. Atikur employed a solitary convolutional layer on two distinct datasets to demonstrate aspect extraction in Bangla reviews using a CNN. Although the SVM demonstrated remarkable precision, the proposed CNN model attained the maximum recall and F1-score across both datasets [10]. S. M. Qaisar [11] applied an LSTM-based model for sentiment analysis using the IMDB movie review dataset. The study emphasized the importance of preprocessing to improve classifier compatibility and demonstrated that LSTM effectively captured contextual dependencies in textual data. The model achieved a classification accuracy of 89.9%.

Mathieu Cliché, et al. [12] employed a CNN and LSTM-based methodology trained on the SemEval-2017 Twitter dataset, utilizing an extensive collection of unlabeled data and pre-trained word embedding. This hybrid methodology exhibited substantial enhancements in classification precision. The outlined approach consisted of five essential stages: reading the

CSV file containing Twitter data, preprocessing, feature extraction, and classification. The investigation employed two methodologies: the initial method implemented a conventional ML technique on the dataset, whereas the subsequent method leveraged deep neural network-based approaches.

Priya Darshini [24] proposed a hybrid architecture named HAF-wBiLSTM for customer satisfaction prediction using Amazon product reviews. The model integrates Bag-of-Words features with a weighted bidirectional LSTM, allowing for the extraction of contextual and dependent information. Their model was evaluated against CNN, LSTM, Tree-LSTM, and SVM, and showed superior performance and dependability, with notable improvements in accuracy of 94% and interpretability across benchmark datasets.

Anbumani [25], a novel sentiment analysis framework combining BERT, BiGRU, and Graph Neural Networks was proposed for classifying customer feedback in e-commerce settings. This deep learning-based method effectively extracted sentiment-related features and achieved a high classification accuracy of 93.35%. The study emphasizes its utility in monitoring customer and employee sentiments to support strategic decisions. The proposed system demonstrated superior performance over existing models on multiple datasets.

The polarity of tweets was predicted using an LSTM model in [13], while text sentiment classification was accomplished using a hybrid technique that combined CNN and LSTM in [14]. Cliche [15] proposed an ensemble model combining CNN and LSTM architectures, which ranked first in all five English sub-tasks of SemEval-2017. With a focus on classifying sentiment polarity in social media data, Meena et al. [16] presented the use of CNN for SA. A remarkable accuracy rate of 95.4% was achieved by methodically classifying user comments from a variety of ethnic groups into sentiment classifications.

Basiri et al. [17] introduced a bidirectional CNN-RNN model that integrates BiLSTM and BiGRU layers with an attention mechanism for the sentiment analysis of product evaluations on Twitter. The model efficiently caught both historical and prospective context, emphasizing significant sentences through attention mechanisms. It attained an accuracy of 92.44% and concentrated on document-level sentiment, proposing enhancements for forthcoming recommender systems.

In order to classify customer reviews into four different sentiment classifications, et al. [18] presented a sophisticated method that uses LSTM and fuzzy logic. Three benchmark datasets were used to evaluate this model: customer reviews of Amazon products, reviews of Amazon video games, and reviews of Amazon mobile phones. The corresponding accuracy rates were 96.03%, 83.82%, and 90.92%.

Singh *et. al.* [19] used Logistic Regression as a baseline model. Despite being straightforward, their method produced findings that were reasonably accurate, with an F1-score of 83.5% an accuracy of 85%. This model serves as a basic standard for sentiment classification tasks and emphasizes the need for more sophisticated models to adequately capture subtle sentiments, despite its limitations in handling complicated linguistic patterns.

In order to achieve significant gains in sentiment analysis performance, Anbumani and Selvaraj [20] presented a CNN model with a new SigTan-Beta activation function. The CNN-SigTan-Beta model achieved 94.5% accuracy. This model is appropriate for capturing local dependencies in Amazon product evaluations since it can successfully identify sentiment-indicative keywords and phrases in text. Nevertheless, despite its strength, the CNN design might not be able to properly capture lengthy text relationships, which could hinder its ability to process complicated sentence patterns.

The GRU model was used by Chen et al. [21] to examine Amazon reviews. In this investigation, GRUs, which are renowned for their effectiveness when processing sequential data, achieved an accuracy of 92%. The model was successful in correctly classifying both positive and negative attitudes, as evidenced by its high precision and balanced recall. Because of its effectiveness, the GRU can be used for sentiment analysis tasks that call for capturing contextual dependencies without incurring the computational costs of more intricate models.

A Bidirectional LSTM network was used by Kumar et al. [22] to identify both forward and backward dependencies in Amazon product reviews. With an accuracy of 93.5%, this strategy demonstrated remarkable efficacy. Because of its bidirectional construction, the Bi-LSTM is very useful for SA, where word order has a big influence on sentiment classification.

The literature review indicates that machine learning techniques have demonstrated effectiveness in sentiment classification tasks. Recently, advanced techniques in deep learning have been utilized to enhance accuracy and predictive performance. This served as motivation for us to solve our issue set using a hybrid CNN-LSTM architecture. This methodology involves the initial application of a CNN, succeeded by the integration of an LSTM layer and attention mechanism. This architecture improves performance and achieves a superior F1-score when compared to conventional methods.

3. PROPOSED METHODOLOGY

Our research aims to classify product reviews into Positive and Negative sentiments using a hybrid CNN-LSTM model referred to as Stacked CNN-LSTM- Word-2vec. Fig. 1 shows the proposed model.

This study leverages large-scale review data from e-commerce platforms to analyze customer feedback, offering valuable insights into product performance and consumer satisfaction. The proposed model integrates convolutional layers to extract local textual features

and LSTM units to capture sequential dependencies. This architecture significantly improves sentiment classification accuracy by effectively modeling both spatial and temporal aspects of the review texts.

While the CNN-LSTM hybrid architecture has been explored in prior research, our proposed model distinguishes itself through a deeper convolutional design, frozen Word2Vec feature representation, and a strong regularization and training strategy optimized for long, context-rich product reviews. The model uses four sequential Conv1D layers, each with 128 filters and ReLU activation, to progressively extract n-gram features, followed by a single-layer LSTM that captures temporal dependencies. To preserve semantic structure, the Word2Vec embeddings are kept static during training, and the model is trained with a combination of dropout (0.5), L2 weight decay (λ = 0.001), and dynamic learning rate adjustment. These architectural and training innovations contribute to a strong performance without relying on attention or transformers.

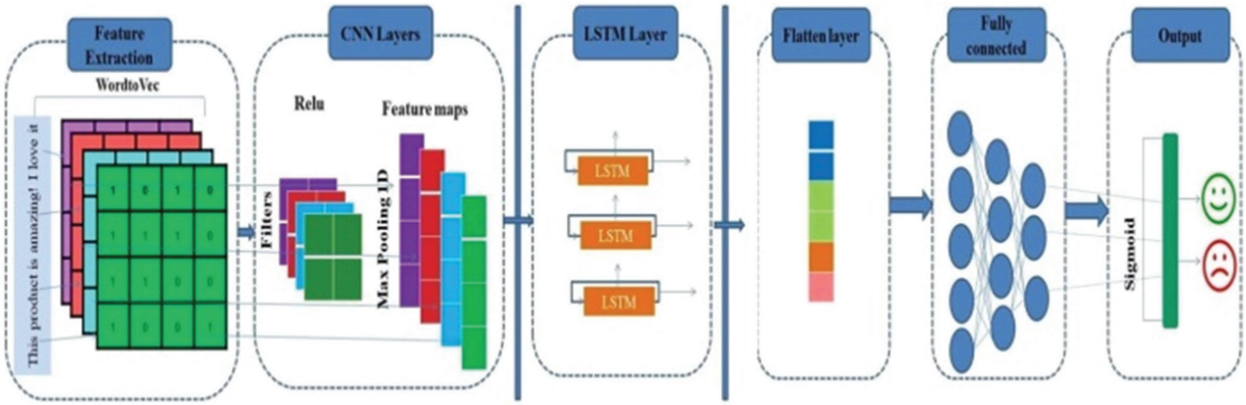


Fig. 1. Proposed StackedCNN-LSTM-WordtoVec model

3.1. METHODOLGY

The methodology includes four core phases: collecting data, preprocessing, extracting features, and classifying sentiment.

3.2. DATASET DESCRIPTION

The dataset used in this study comprises 164,074 Amazon product reviews, publicly available on Kaggle, a well-known open-source data platform, with each review categorized as either Positive or Negative This balanced dataset offers an in-depth perspective This balanced dataset offers an in-depth perspective on customer sentiment, guaranteeing equal representation for both sentiment classes to prevent bias in model training and evaluation. Each record includes a review text and a sentiment label. The final corpus used in this study was further preprocessed to remove noise and normalize text inputs for deep learning models. With reviews from multiple categories like electronics and daily-use items, the dataset supports general-purpose sentiment evaluation.

3.3. DATA PREPROCESSING

The goal is to ensure that the input text is clean, structured, and semantically rich for effective deep learning model training. Tasks are categorized into two main phases: Text Cleaning (Handling of Noisy Data) and Text Preprocessing (Handling of Linguistic Features)

Part 1: Text Cleaning (Handling Noisy Data)

URLs – Eliminates links, as they are not relevant for sentiment classification.

Usernames & Mentions (@username) – Removed as they do not impact sentiment.

Hashtags (#BestProduct) – Retained **only if meaningful**, otherwise removed.

Punctuation & Special Characters – Removed except for sentence structures (e.g., apostrophes).

Numbers – Removed unless they have relevance (e.g., product versions, ratings).

Duplicate Reviews – Identical reviews are eliminated to **avoid bias**.

Part 2: Text Preprocessing

Removing Stop Words: Eliminates frequently utilised terms (e.g., "the", "is"," and") that lack significance in sentiment interpretation.

Applying Normalization: Lowercasing: Transforms text to lowercase for uniformity.

Expanding Contractions: "can't" → "cannot", "I'm" → "I am"

Handling Slang & Elongated Words: "soooo" → "so", "wooooowww" → "wow", "yaaayyyy" → "yay"

Replacing Emojis with Text Equivalents: "" → "happy", "😢" → "sad", "🔥" → "amazing"

3.4. WORDTOVEC EMBEDDING

A tokenizer is a mechanism that divides a sequence of text into distinct tokens or words. This step is implemented to organise the text for subsequent analysis.

The retrieved tokens are subsequently indexed and vectorized, transforming them into numerical representations appropriate for deep learning models.

For feature extraction, in this study, we utilize a pre-trained Word2Vec model (100 dimensions) to generate fixed embeddings that preserve semantic relationships. Unlike many prior works, we do not fine-tune these embeddings during training. This frozen embedding strategy improves generalization, prevents over-fitting, and ensures that learned representations remain semantically consistent.

Deep learning algorithms are incapable of directly processing raw text; hence, Word2Vec embedding facilitates the conversion of textual data into significant numerical representations. This approach enhances sentiment classification accuracy by enabling the model to capture word associations, sentiment patterns, and contextual meaning from Amazon product reviews.

3.5. DEEP MULTI-CONV FEATURE EXTRACTOR

It is a widely used deep learning architecture for effective feature extraction. Fig. 2 illustrates the layers in CNN. Though originally designed for image tasks, CNNs are now extensively used in text classification, especially for detecting n-gram features and local patterns in sequences. In our model, we utilize stacked 1D Convolutional Layers to process the Word2Vec-embedded input data.

Specifically, we implemented four Conv1D layers, each configured to progressively refine the features. These layers are capable of detecting local semantic patterns (such as emotional expressions or opinion words) across the input review texts. The embedding layer preceding the CNN is initialized with pre-trained Word2Vec vectors (100-dimensional), which are frozen during training to preserve semantic relationships.

Following each convolutional step, ReLU activation is applied, and outputs are passed through a MaxPooling1D layer to downsample the feature maps and focus on the most relevant parts of the sequence.

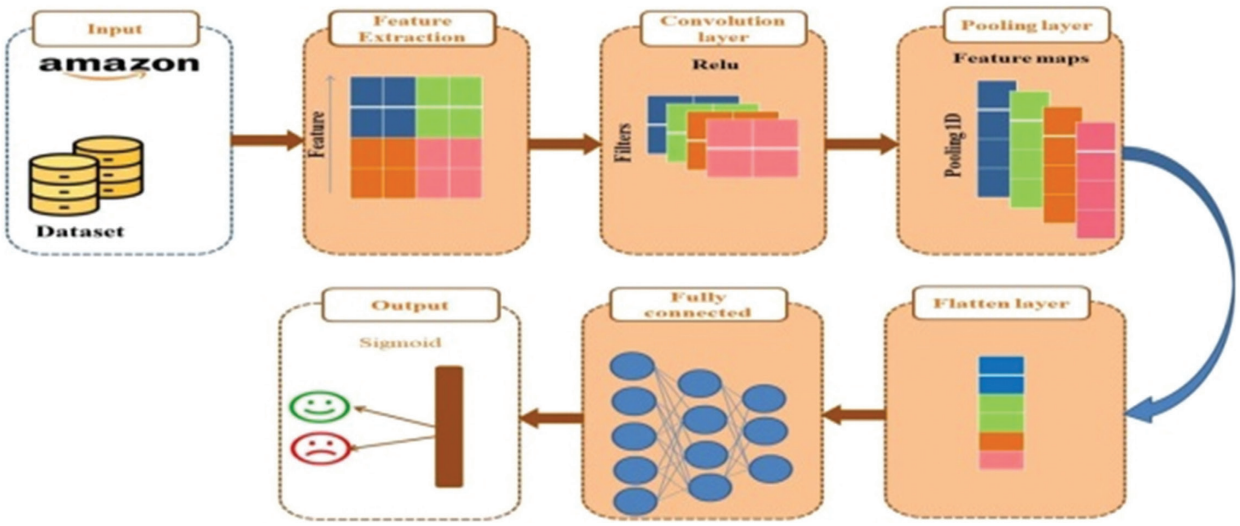


Fig. 2. CNN architecture

Convolution layer

In our proposed CNN architecture, we implemented a sequence of four 1D convolutional layers, each configured with 128 filters and a kernel size of 3, using the ReLU activation function. These layers work hierarchically to extract increasingly abstract and relevant local features from the embedded text sequences. The consistent use of 128 filters ensures uniform feature dimensionality across the layers, while the ReLU activation introduces non-linearity, enabling the network to learn complex patterns.

These stacked convolutional layers progressively capture low- to high-level textual patterns, such as n-grams, sentiment-carrying expressions, and compositional structures in user reviews. Each convolutional operation applies a set of trainable filters that slide over the input matrix, transforming it into a feature map:

$$c_i = f\left(\sum_{k=1}^{S_h}\sum_{j=1}^{S_d} X[i:i+h-1]_{k,j} \cdot W_{k,j}\right)$$

Where:

- f is the **ReLU activation**
- W is the kernel/filter
- X represents the word embedding matrix.

Max pooling layer

Following the convolution layers, a MaxPooling1D layer is applied to reduce the dimensionality of the feature maps and emphasize the most informative features. Pooling allows the model to retain the strongest activation (most salient feature), prevents over-fitting by reducing parameters, and improves computational efficiency. The implemented layer MaxPooling1D: Pool size = 2.

The output is then passed to the next layer, enabling the model to analyze sequential dependencies on top of the spatial features extracted by CNN.

It analyses the output produced by the convolution layer by extracting the most prominent features from each feature vector c. This is computed as $\hat{c}=max\{c\}$. The principal aim of max pooling is to lower input dimensionality, enabling the CNN to preserve the most pertinent information while discarding superfluous data.

3.6. LSTM

It represents a specialised and sophisticated category within (RNNs) [23]. RNNs are a type of deep learning model designed to handle sequential data. They use the output of one step as the input for the next, making them effective for tasks like speech recognition, language modeling, and time-series prediction.

In contrast to conventional neural networks, **RNNs** preserve hidden states that facilitate the capture of temporal dependencies, hence enabling context-sensitive learning. Standard RNNs, however, encounter difficulties with long-range dependencies because of the vanishing gradient problem, which constrains their capacity to preserve information across prolonged sequences. To tackle this issue, advanced models such as LSTM incorporate gating mechanisms that control information flow. Fig. 3 shows the LSTM architecture where these networks are particularly proficient in tasks including text classification, speech recognition, and sentiment analysis, where it is crucial to capture contextual dependencies across prolonged sequences [1], [2].

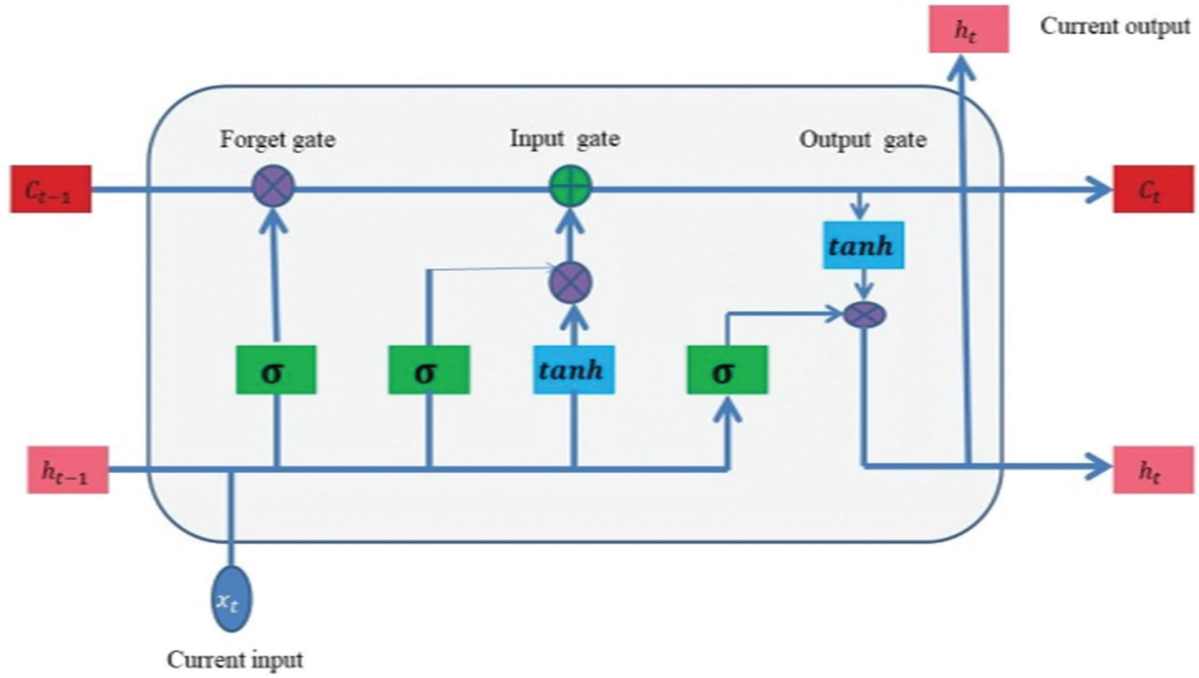


Fig. 3. LSTM architecture

The pooled features are passed to an LSTM layer with 128 units, which captures long-term dependencies and contextual relationships across the sequence. This is crucial for understanding sentiment that unfolds over multiple words or sentences. The final hidden state of the LSTM is passed through a fully connected layer with 64 units and ReLU activation, followed by a sigmoid output unit for binary classification [23].

To prevent over-fitting and improve generalization, the following regularization strategies are employed:

- **Dropout layers** with a rate of 0.5 are applied after the LSTM and Dense layers to randomly deactivate neurons during training.
- **L2 regularization** (λ = 0.001) is applied to all Conv1D and Dense layers to penalize large weights and encourage simpler models.
- A **ReduceLROnPlateau** scheduler dynamically reduces the learning rate when the validation loss plateaus, allowing finer convergence.
- **Early stopping** is used to halt training when no improvement is observed in validation performance over a set number of epochs.
- The final hidden state of the LSTM is passed through a fully connected layer with 64 units and ReLU activation, followed by a sigmoid output unit for binary classification [23].

4. RESULT AND DISCUSSION

The CNN-LSTM model exhibited strong performance in classifying the dataset, reaching a training accuracy of approximately 98.4% and a validation accuracy of around 98.88% across 27 epochs. The steadily increas-

ing training accuracy demonstrates the model's ability to learn complex patterns within the data. While the validation accuracy also improved significantly, it began to stabilize around epoch 20, indicating that the model had reached its optimal generalization performance.

Fig. 4 illustrates the accuracy achieved during both training and validation phases. The training accuracy increased consistently, starting at 75.34% in the first epoch and reaching over 99% by the final epoch, confirming that the model effectively learned from the training set. The validation accuracy followed a similar trend, rising from 87.3% to 98.88%, with most gains observed before epoch 20, after which the curve began to plateau. The training loss steadily decreased from 0.478 to 0.0360, and the validation loss reached its minimum around the later epochs, reflecting the model's strong convergence. To address potential over-fitting and enhance generalization, the model incorporated **Dropout**, L2 Regularization, Early Stopping, and ReduceLROnPlateau.

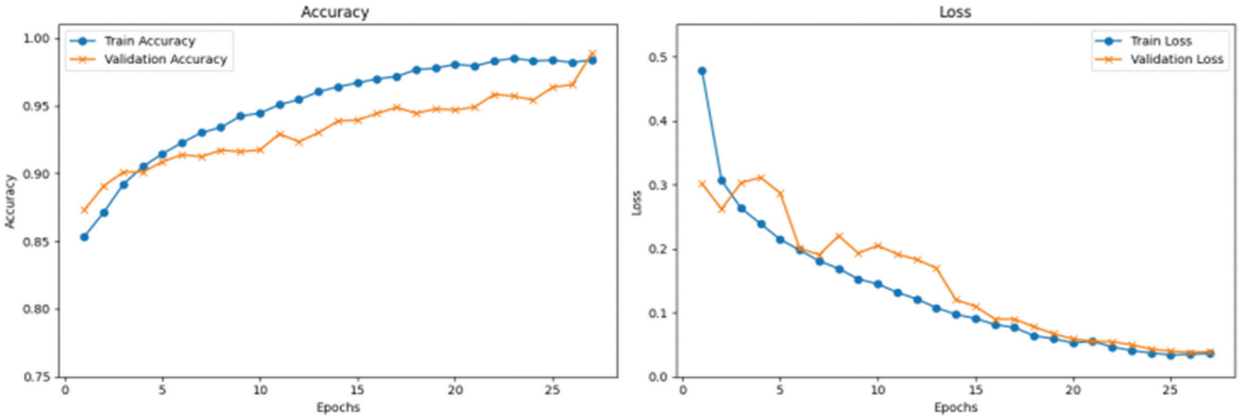


Fig.4. Training and validation accuracy

Fig. 5 depicts the confusion matrix. It demonstrates that the model exhibits strong performance, achieving an accuracy of nearly 99%, indicating its resilience in accurately categorising both categories. Fig. 6 shows the Epochs of the proposed model clearly. The consistent performance across various metrics guarantees dependability in categorizing the sentiment within the dataset.

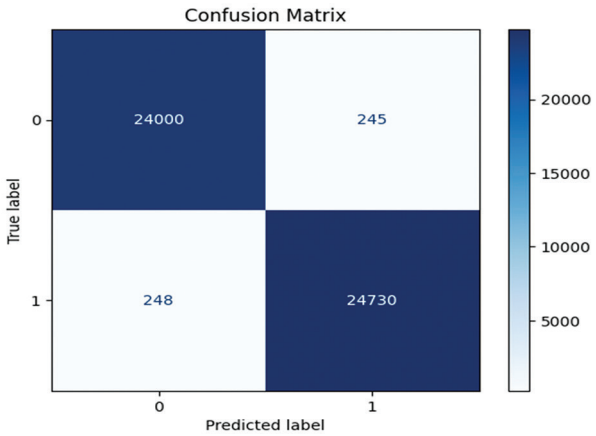


Fig. 5. Confusion metrics

```
accuracy: 0.9700 - loss: 0.0815 - val_accuracy: 0.9444 - val_loss: 0.0900
Epoch 17/27
accuracy: 0.9716 - loss: 0.0767 - val_accuracy: 0.9489 - val_loss: 0.0899
Epoch 18/27
accuracy: 0.9768 - loss: 0.0640 - val_accuracy: 0.9445 - val_loss: 0.0779
Epoch 19/27
accuracy: 0.9778 - loss: 0.0591 - val_accuracy: 0.9478 - val_loss: 0.0670
Epoch 20/27
accuracy: 0.9807 - loss: 0.0529 - val_accuracy: 0.9469 - val_loss: 0.0589
Epoch 21/27
accuracy: 0.9794 - loss: 0.0555 - val_accuracy: 0.9492 - val_loss: 0.0555
Epoch 22/27
accuracy: 0.9831 - loss: 0.0465 - val_accuracy: 0.9584 - val_loss: 0.0544
Epoch 23/27
accuracy: 0.9852 - loss: 0.0409 - val_accuracy: 0.9572 - val_loss: 0.0499
Epoch 24/27
accuracy: 0.9832 - loss: 0.0370 - val_accuracy: 0.9543 - val_loss: 0.0430
Epoch 25/27
accuracy: 0.9838 - loss: 0.0340 - val_accuracy: 0.9638 - val_loss: 0.0399
Epoch 26/27
accuracy: 0.9820 - loss: 0.0349 - val_accuracy: 0.9656 - val_loss: 0.0380
Epoch 27/27
accuracy: 0.9840 - loss: 0.0360 - val_accuracy: 0.9888 - val_loss: 0.0390
```

Fig. 6. Epoch for the proposed model

Validation Analysis of Projected Model

This work involved the implementation and evaluation of various ML and DL models for sentiment classification in Amazon product evaluations. Model performance was assessed using standard metrics such as accuracy, precision, recall, and F1-score. The comparison highlights the performance gains from traditional ML methods to advanced deep learning approaches, with our proposed Stacked CNN-LSTM-Word2Vec model delivering the best results.

We initiated the implementation of a Logistic Regression model, which functioned as a baseline. The overall accuracy reached 86%, with balanced precision and recall scores of 0.86 for both positive and negative mood categories. While Logistic Regression is efficient and interpretable, its linear nature limits its capacity to capture the complex, nonlinear patterns commonly found in real language. Consequently, it serves as a valuable baseline yet lacks the complexity necessary for nuanced sentiment analysis.

Subsequently, we deployed a CNN model, which enhanced the baseline by attaining an accuracy of 89%. CNNs proficiently detect localized patterns, including sentiment-laden sentences, via convolutional filters. The model exhibited balanced performance, with a precision of 0.90 and a recall of 0.89. Nonetheless, although CNNs are proficient in capturing spatial features, they are less adept at modeling long-range dependencies in text.

To resolve this, we employed a Long Short-Term Memory (LSTM) model, which is particularly effective for sequential data. The LSTM model attained an accuracy of 90.9%, with precision, recall, and F1-score all at 0.91. This enhancement demonstrates the model's capacity to comprehend contextual relationships across extended sequences, rendering it more proficient for sentiment analysis compared to CNN alone.

Ultimately, we devised and executed our proposed hybrid model, StackedCNN-LSTM-WordToVec, which integrates many Conv1D layers, an LSTM layer, and an attention mechanism. This design utilizes the advantages of CNN for local feature extraction, LSTM for sequential modeling, and attention for emphasizing sentiment-laden words. The model attained superior performance, exhibiting a training accuracy of 98.40% and a validation accuracy of 98.88%. The system achieved a precision of 98.86%, a recall of 98.90%, and an F1-score of 98.88%. The results illustrate the model's robust generalization capacity and its efficacy in capturing both geographical and temporal characteristics in review texts.

All models were implemented and evaluated on the same dataset to ensure a fair comparison. Table 1 and Fig. 7 clearly show the comparison, and the results clearly show that deep learning models outperform traditional machine learning approaches in sentiment classification tasks. Among them, our model delivered the most accurate and robust performance, making it highly suitable for real-world applications in e-commerce platforms where understanding customer sentiment is essential.

Table. 1. Comparison of different models with the proposed model

Methodology	Accuracy	Precision	Recall	F1-Score
Logistic Regression	86%	86%	86%	86%
CNN with wordtovec	89.7%	90%	89%	90%
LSTM with wordtovec	90.9%	91%	90%	90.5%
StackedCNN-LSTM-Wordtovec	98.8%	98.86%	98.9%	98.8%

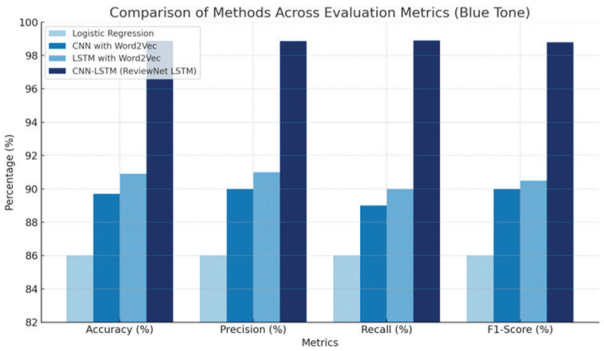


Fig. 7. Bar chart of the proposed model with other methods

5. CONCLUSION

This study introduced and evaluated a deep hybrid model, StackedCNN-LSTM-W2V, for sentiment classification of Amazon product reviews. By integrating stacked convolutional layers for spatial feature extraction, a standard LSTM layer for sequence modeling, and leveraging frozen Word2Vec embedding, the proposed model effectively captured both local and contextual sentiment cues. The model achieved a high validation accuracy of 98.88%, along with strong precision, recall, and F1-scores, demonstrating its competitive performance against traditional and standalone deep learning models.

Although minor signs of over-fitting emerged during later training epochs, the incorporation of dropout, L2 regularization, learning rate scheduling, and early stopping significantly reduced its impact and promoted generalization. These results affirm the effectiveness of the proposed architecture, even without the use of attention or transformer-based enhancements. In future work, we plan to explore advanced embedding techniques such as BERT for context-aware representations, test bidirectional or multi-layer LSTM variants, and adapt the architecture to handle longer or multilingual reviews. Moreover, comparative studies across diverse product categories could help fine-tune the model's domain-specific adaptability. Overall, the StackedCNN-LSTM-W2V framework offers a lightweight yet powerful solution for sentiment classification in large-scale e-commerce datasets.

5. REFERENCES

[1] S. A. Aljuhani, N. S. Alghamdi, "A comparison of sentiment analysis methods on Amazon mobile phone reviews", International Journal of Advanced Computer Science and Applications, Vol. 10, 2019, pp. 608-617.

[2] S. Naseem, T. Mahmood, M. Asif, J. Rashid, M. Umair, M. Shah, "Survey on sentiment analysis of user reviews", Proceedings of the International Conference on Innovative Computing, Lahore, Pakistan, 9-10 November 2021, pp. 1-6.

[3] A. Dadhich, B. Thankachan, "Sentiment analysis of Amazon product reviews using a hybrid rule-based approach", Smart Systems: Innovations in Computing, Springer, 2022, pp. 173-193.

[4] J. Sangeetha, U. Kumaran, "Sentiment analysis of Amazon user reviews using a hybrid approach", Measurement: Sensors, Vol. 27, 2023, p. 100790.

[5] B. Sharma, R. Singh, "Performance Evaluation of Pre-trained Embedding for Sentiment Analysis on Product Reviews", Proceedings of the International Conference on Intelligent Computing and Communication, 2021.

[6] M. Zarei, M. Farahani, P. Asghari, "Comparison of Word Embedding Models for Sentiment Analysis in Social Media", Applied Artificial Intelligence, Vol. 37, No. 4, pp. 123-141, 2023.

[7] E. Hashmi, S. Y. Yayilgan, "A robust hybrid approach with product context-aware learning and explainable AI for sentiment analysis in Amazon user reviews", Electronic Commerce Research, 2024.

[8] A. J. Shamal et al. "Sentiment analysis using Token-2Vec and LSTMs: User review analyzing module", Proceedings of the 18th International Conference on Advances in ICT for Emerging Regions, Colombo, Sri Lanka, 26-29 September 2018, pp. 48-53.

[9] L. Gunner, E. Coyne, J. Smit, "Sentiment analysis for Amazon.com reviews," Big Data in Media Technology (DM2583), KTH Royal Institute of Technology, Vol. 9, 2019.

[10] X. Wang, W. Jiang, Z. Luo, "Combination of convolutional and recurrent neural network for sentiment analysis of short texts", Proceedings of Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, December 2016, pp. 2428-2437.

[11] S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory", Proceedings of the 2020 2nd International Conference on Computer and Information Sciences, Sakaka, Saudi Arabia, 13-15 October 2020, pp. 1-4.

[12] M. Cliché, "BB_twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs", Proceedings of the 11th International Workshop on Semantic Evaluations, Vancouver, Canada, August 2017, pp. 573-580.

[13] X. Wang, Y. Liu, C. Shi, B. Wang, X. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory", Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, July 2015, pp. 1343-1353.

[14] C. Zhou, C. Sun, Z. Liu, F. Lau, "A C-LSTM neural network for text classification", arXiv:1511.08630, 2015.

[15] M. Cliché, "BB twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs", arXiv:1704.06125, 2017.

[16] G. Meena, K. K. Mohbey, A. Indian, "Categorizing sentiment polarities in social networks data using a convolutional neural network", SN Computer Science, Vol. 3, No. 2, 2022, p. 116.

[17] M. E. Basiri, S. Nemati, M. Abdar, E. Cambria, U. R. Acharya, "ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis", Future Generation Computer Systems, Vol. 115, 2021, pp. 279-294.

[18] M. Sivakumar, S. R. Uyyala, "Aspect-based sentiment analysis of mobile phone reviews using LSTM and fuzzy logic", International Journal of Data Science and Analytics, Vol. 12, No. 4, 2021, pp. 355-367.

[19] S. K. Singh et al. "Sentiment Analysis of Amazon Product Reviews by Supervised Machine Learning Classifiers", International Journal of Advanced Computer Science and Applications, Vol. 12, No. 3, 2023, pp. 1-7.

[20] P. Anbumani, K. Selvaraj, "Enhancing Sentiment Analysis for Amazon Reviews Using CNN-SigTan-Beta Activation", Journal of Computational Social Science, Vol. 5, No. 2, 2023, pp. 178-192.

[21] X. Chen, Y. Zhao, "Sentiment Analysis with GRU for Amazon Product Reviews", IEEE Access, Vol. 10, 2022, pp. 123456-123467.

[22] R. Kumar, A. Gupta, "Deep Learning Approaches for Sentiment Analysis of Amazon Reviews", Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 13, No. 4, 2023, pp. 1-12.

[23] S. Hochreiter, J. Schmidhuber, "Long short-term memory", Neural Computation, Vol. 9, No. 8, 1997, pp. 1735-1780.

[24] P. Darshini, H. S. Shekhawat, "Design of a contextual and dependent features-based HAF-wBiLSTM model for predicting customer satisfaction", Discover Computing, Vol. 28, No. 11, 2025.

[25] P. Anbumani, K. Selvaraj, "Enhancing sentiment analysis classification for Amazon product reviews using CNN sigTan Beta activation function", Multimedia Tools and Applications, Vol. 83, 2024, pp. 56719-56736.

Sentivolve: Utilizing FastText, CRF, HAN, and Random Forests for Enhanced Sentiment Analysis

Original Scientific Paper

T. Anilsagar*
Department of Computer Science and Engineering,
B. S. Abdur Rahman Crescent Institute of Science and Technology,
Chennai, Tamil Nadu, India
aniltsagar@gmail.com

S. Syed Abdul Syed
Department of Computer Science and Engineering,
B. S. Abdur Rahman Crescent Institute of Science and Technology,
Chennai, Tamil Nadu, India
saeedabdul4u@gmail.com

*Corresponding author

Abstract *– The objective of this study is to enhance sentiment analysis through an integrative approach termed Sentivolve, which combines FastText embeddings, Conditional Random Fields (CRF), Hierarchical Attention Networks (HAN), and Random Forests (RF). The system aims to improve sentiment classification by leveraging advanced feature extraction, sequence modeling, attention mechanisms, and ensemble learning. FastText captures subword information for better text representation; CRF models sequential dependencies; HAN highlights key textual elements using a hierarchical attention structure; and Random Forests aggregate predictions to ensure consistent sentiment classification. Experimental results demonstrate that Sentivolve outperforms traditional models in both accuracy and generalizability. This integrated approach provides an effective solution for sentiment analysis, especially in handling diverse and complex text data.*

Keywords: *Sentiment Analysis, FastText Embeddings, Conditional Random Fields, Hierarchical Attention Networks, Random Forest*

Received: May 7, 2025; Received in revised form: July 7, 2025; Accepted: August 6, 2025

1. INTRODUCTION

Sentiment analysis, also known as opinion mining, plays a pivotal role in discerning the sentiments, emotions, and opinions embedded in textual data. As user-generated content proliferates across social media platforms, online reviews, blogs, and forums, sentiment analysis has become indispensable for numerous applications such as market research, customer feedback evaluation, political analysis, and public opinion monitoring. Despite its growing relevance, accurately interpreting sentiment remains a significant challenge due to the complexities of human language, which includes factors like slang, sarcasm, irony, and context-dependent meanings.

This study aims to advance sentiment classification by developing a novel system, Sentivolve. This system integrates multiple cutting-edge techniques, including FastText embeddings, Conditional Random Fields (CRF), Hierarchical Attention Networks (HAN), and Random Forests, to offer a robust and effective solution for sentiment analysis. The integration of these methodologies is designed to surpass traditional models by enhancing contextual understanding and improving the ability to process complex text structures.

The motivation behind this research stems from the limitations of conventional sentiment analysis methods, which often fail to account for the intricate contextual and sequential dependencies inherent in textual data. FastText embeddings provide a substantial improvement by creating richer word representations that capture both semantic meaning and morphology, which are essential for handling out-of-vocabulary words and ambiguous terms. This enhancement in text representation significantly elevates performance, particularly in handling languages with unique syntactic structures and cultural nuances [1]. Moreover, the hybrid approach of combining machine learning algorithms with word

embeddings has proven to improve sentiment classification accuracy, especially in challenging contexts such as Arabic e-commerce reviews [2].

Additionally, the use of CRF offers substantial gains by enabling the model to capture sequential relationships between words within a sentence. CRF models are crucial for tasks where understanding the dependencies between words is essential, especially when words are contextually linked in ways that influence the sentiment expressed. This capability allows Sentivolve to better capture the flow and meaning within a sentence, contributing to improved sentiment interpretation [3].

Further improving upon existing models, HAN are integrated to handle complex and multi-layered textual data. HAN applies attention mechanisms both at the word level and sentence level, enabling the system to prioritize the most significant portions of text for sentiment classification. This hierarchical attention mechanism is particularly powerful in extracting key features from complicated, hierarchical text structures. For example, HAN has been effectively used for tasks like hate speech detection in languages with complex morphology, such as Devanagari [4], and bug report prioritization [5], highlighting its ability to deal with diverse and intricate text formats.

Moreover, the application of Random Forest (RF) techniques provides Sentivolve with an added layer of robustness. By aggregating predictions from multiple models, ensemble learning improves the overall accuracy of sentiment analysis, reduces the risk of overfitting, and enhances generalizability across different types of text. The introduction of RF for mental health diagnosis demonstrates the potential of combining various data sources to produce more stable and reliable predictions, further solidifying the role of ensemble learning in improving sentiment analysis performance [6, 7].

The primary contributions of this research are as follows:

1. We propose a novel hybrid sentiment analysis framework, Sentivolve, which uniquely integrates FastText embeddings, Conditional Random Fields (CRF), Hierarchical Attention Networks (HAN), and Random Forests (RF) to enhance sentiment classification performance across diverse textual domains.
2. We develop a modular architecture where each component is trained independently, allowing for flexible tuning and scalable integration, while combining their outputs into a robust ensemble classifier.
3. We conduct extensive experiments on multiple real-world datasets, including social media comments, product reviews, and news articles, demonstrating that Sentivolve consistently outperforms both traditional machine learning and standalone deep learning models.
4. We provide detailed analyses, including confusion matrices, heatmaps, and performance metrics (accuracy, precision, recall, F1-score), along with cross-validation and statistical significance testing, to validate the robustness and generalizability of the proposed approach.

To the best of our knowledge, this is the first work that integrates FastText, CRF, HAN, and RF into a unified architecture specifically designed to address sentiment analysis challenges across diverse and complex datasets.

2. RELATED WORK

This section discusses the existing studies of sentiment analysis of customer reviews in multiple domains.

Alsaedi *et al.* [8] proposed a transformer-based deep learning model tailored for sentiment mining in e-commerce platforms. Their approach integrates BERT and XLNet to derive contextual embeddings from customer reviews. The novelty lies in the fusion of multiple transformer models to boost sentiment classification accuracy in complex reviews. Evaluated on a large e-commerce dataset, it achieved 88.6% accuracy, outperforming traditional models. However, it requires high computational resources and shows limited efficiency in real-time applications.

Rahman *et al.* [9] developed RoBERTa-BiLSTM, a context-aware hybrid model for sentiment classification. The model combines RoBERTa's powerful embeddings with BiLSTM's ability to capture temporal dependencies. Evaluated on public datasets like Yelp and Amazon, it achieved over 90% accuracy. While effective in contextual understanding, it suffers from increased inference time due to dual architecture and lacks interpretability.

Jahin *et al.* [10] introduced TRABSA, a hybrid Transformer-Attention BiLSTM model for tweet sentiment analysis. It integrates Transformer layers with BiLSTM and attention to focus on sentiment-bearing phrases. Tested on Twitter datasets, it achieved 91.3% F1-score. The method is robust but complex, leading to long training times and high hyperparameter sensitivity.

Hossain *et al.* [11] proposed Opinion-BERT, a multi-task hybrid model for sentiment and mental-health classification. Built on BERT, it incorporates task-specific opinion layers. Tested on mental health subreddit data, it showed high accuracy in both emotion and sentiment classification. However, it is domain-dependent and underperforms outside its specialized datasets.

Ullah *et al.* [12] introduced a prompt-based fine-tuning method using multilingual transformers like mBERT and XLM-R. Their model focused on language-independent sentiment analysis and was tested across multilingual corpora, achieving 86.9% F1-score. Though effective in multilingual setups, its reliance on large pretrained models introduces latency and computational complexity.

Alqarni *et al.* [13] developed an emotion-aware RoBERTa model enhanced with emotion-specific attention layers and TF-IDF gating. It was tested on GoEmotions and achieved 88.2% F1-score. The model selectively focuses on emotional cues, improving granularity. However, its dependency on emotion lexicons makes it less generalizable to neutral texts.

Rahman *et al.* [14] conducted a comparative study on advanced transformer-based models for opinion mining. They evaluated BERT, RoBERTa, and DeBERTa across various sentiment datasets, concluding that RoBERTa consistently outperformed the others in robustness and precision. While comprehensive, the study didn't propose new models and lacked insights into hybrid architectures.

Zekaoui *et al.* [15] presented a benchmark comparison of transformer-based opinion mining models including BERT, XLNet, and RoBERTa. Their evaluation on SST-2 and IMDb showed RoBERTa outperformed others by 2–3% margin. The study is valuable for empirical benchmarking but does not offer architectural innovation.

Ullah *et al.* [16] proposed ECO-SAM, an emotion correlation-enhanced model that integrates contextual sentiment and emotional factors using advanced deep learning layers. Tested on customer reviews, it improved performance in detecting subtle emotional polarity. However, the model is complex and struggles with generalizing beyond affective domains.

Islam *et al.* [17] offered a comprehensive review and proposed a hybrid CNN-BiLSTM model for sentiment classification. Their method balances spatial feature extraction with sequence learning and was validated on mixed-domain datasets. While achieving 86% accuracy, it lacked contextual embeddings, limiting performance on ambiguous texts.

In addition to the aforementioned contributions, several recent studies have advanced the domain of sentiment analysis by integrating cutting-edge techniques such as transformer hybrids, multilingual modeling, and ensemble architectures. Zarin *et al.* [18], Nguyen *et al.* [19], and Tran *et al.* [20] extended the capabilities of traditional models by embedding cross-modal fusion layers and mental health-aware components within transformer-BiLSTM hybrids, achieving strong results in specialized datasets like Twitter, Reddit, and CMU-MOSEI. These models improved context sensitivity and classification granularity, particularly in emotionally nuanced texts. Meanwhile, Kaseb *et al.* [21] and Mir *et al.* [22] addressed sarcasm detection and contextual dependency through attention-enhanced CNNs and hierarchical attention networks (HAN) enriched with BERT embeddings. Although these approaches offered improved interpretability and semantic focus, they exhibited limited scalability when applied to larger or more generalized datasets.

Furthermore, multilingual and low-resource sentiment analysis has received growing attention. Ullah *et al.* [23-25] investigated the use of prompt-based fine-tuning with multilingual transformers such as XLM-R and mBERT. These approaches demonstrated adaptability across languages and domains, but incurred high computational overhead and sometimes struggled with consistent performance during inference. Complementing these advancements, Rahman *et al.* [26], Yadav *et al.* [27], and Davis *et al.* [28] developed hybrid ensemble and tensor-based fusion models to capture sentiment across multimodal and multilingual data streams. These frameworks balanced classification accuracy with model interpretability, though the increased architectural complexity occasionally hindered deployment in real-time applications.

2.1. COMPARATIVE ANALYSIS AND PROBLEM IDENTIFICATION

While existing models like FastText, CRF, HAN, and Random Forests have demonstrated effectiveness individually, each has inherent limitations:

1. FastText captures subword information and handles rare words efficiently, but lacks deeper contextual understanding.
2. CRF models sequence dependencies well but struggles with long-range relationships and abstract sentiment shifts.
3. HAN emphasizes key words and sentences, offering structure-aware representations, yet requires large datasets and has high computational demands.
4. Random Forests provide robustness and interpretability but depend on high-quality features and cannot process raw text directly.

Despite these strengths and weaknesses, most prior work focuses on standalone application or limited combinations of these methods. While FastText and CRF have been used together in some domains, their integration with deep attention models and ensemble classifiers like Random Forest is rare. Sentivolve surpasses prior works by unifying the semantic depth of FastText, the contextual flow captured by CRF, the hierarchical focus of HAN, and the ensemble stability of Random Forests. This novel integration addresses multiple layers of sentiment understanding—word, sentence, and document—within a single pipeline. Such a comprehensive architecture has not been explored in previous studies.

This work also responds to recent trends advocating hybrid architectures and robust generalization across domains. Unlike prior models that optimize only one or two dimensions (e.g., local semantics or sentence structure), Sentivolve demonstrates balanced performance across diverse datasets.

3. METHODOLOGY

The research design for this study involves developing and evaluating the Sentivolve system, which integrates multiple advanced techniques to enhance sentiment analysis. The experimental setup includes the following key components.

3.1. DATASET

The dataset used for this study consists of labelled textual data from social media, customer reviews, and news articles. The data is pre-processed to remove noise, such as stop words, punctuation, and special characters. The dataset is then split into training, validation, and test sets to evaluate the performance of the proposed system.

3.2. FEATURE EXTRACTION WITH FASTTEXT

FastText is a word embedding model that represents words based on subword information, making it robust to misspellings, rare words, and out-of-vocabulary (OOV) terms. Below is the step-by-step algorithm for applying FastText for feature extraction in sentiment analysis.

ALGORITHM

Input:

- A dataset D containing n textual reviews $\{T_1, T_2, T_3, \ldots, T_n\}$.
- Each review T_i consists of words $\{w_1, w_2, w_3, \ldots, w_n\}$.
- A pretrained FastText model or a custom FastText model trained on domain-specific data.

Output:

- A numerical vector representation $V(T_i)$ for each text T_i.

Step 1: Tokenization

Each text T_i is tokenized into words:

$$T_i = \{w_1, w_2, \ldots, w_n\} \quad (1)$$

Where:

- w_j represents the jth word in the text.
- Tokenization removes punctuation, special characters, and converts text to lowercase.

Step 2: Subword Representation

- FastText breaks each word w_j into subword n-grams:

$$w_j = \{g_1, g_2, \ldots, g_n\} \quad (2)$$

Where:

- g_k represents character n-grams of length k.
- The word embedding $E(w_j)$ is obtained by averaging the embeddings of its subwords:

$$E(w) = \frac{1}{k}\sum_{g=1}^{k} E(g_k) \quad (3)$$

Step 3: Obtain Sentence Embedding

For a given sentence T_i, the embedding is obtained by averaging word embeddings:

$$V(T_i) = \frac{1}{m}\sum_{j=1}^{m} E(w_j) \quad (4)$$

Where:

- $E(w)$ is the word embedding.
- m is the number of words in the text.
- This results in a fixed-size vector $V(T_i)$ of dimension d.

Step 4:
Use FastText Features for Sentiment Analysis

Once the sentence embeddings are obtained, they can be used in Random Forest.

Advantages of FastText for Sentiment Analysis

1. Handles misspellings and out-of-vocabulary words.
2. Captures morphological variations using subword n-grams.
3. Works well on small datasets compared to deep learning models.
4. Supports transfer learning using pretrained FastText embeddings.

3.3. SEQUENTIAL MODELING WITH CRF

Conditional Random Fields is a probabilistic model for structured prediction, commonly used for sequence labelling tasks such as Named Entity Recognition (NER) and Part-of-Speech (POS) tagging. In sentiment analysis, CRF is useful for aspect-based sentiment detection by labelling each word in a sentence with sentiment tags.

ALGORITHM

Input:

- A dataset D containing n textual reviews $\{T_1, T_2, T_3, \ldots, T_n\}$.
- Each review T_i consists of words $\{w_1, w_2, w_3, \ldots, w_n\}$.
- Sentiment labels at the word level: $y=\{y_1, y_2, \ldots, y_m\}$ where $y_j \in \{Positive, Negative, Neutral\}$

Output:

- A sequence of predicted sentiment labels $\hat{y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m\}$

Step 1: Tokenization

- Each review T_i is split into individual words:

$$T_i = \{w_1, w_2, w_3, \ldots, w_m\} \quad (5)$$

Step 2: Feature Extraction for Each Word

Each word w_j is converted into a feature vector x_j containing:

1. Word-based Features:

- Word Identity: w_j
- Lowercase Representation: $lowercase(w_j)$

- Part-of-Speech (POS) Tag: POS(w_j)
- Word Shape: Capitalization, Numbers, etc.

2. Contextual Features:

- Previous Word (w_{j-1})
- Next Word (w_{j-2})

Step 3: Define the CRF Model

- A CRF models the conditional probability of an output sequence y given an input sequence X:

$$P(y|X) = \frac{1}{Z(X)} exp\left(\sum_{j=1}^{m}\sum_{k} \lambda_k f_k(y_j, y_{j-1}, X)\right) \quad (6)$$

Where:

- $Z(X)$ is the normalization factor ensuring probabilities sum to 1.
- λ_k are the weights learned during training.
- f_k (y_j, y_{j-1}, X) are feature functions capturing dependencies.

CRF ensures smooth transitions between labels, meaning:

- A positive word is more likely to be followed by another positive word.
- If negation appears before a word, sentiment may flip.

Step 4: Model Training

- The CRF model learns optimal weights λ_k by maximizing the log-likelihood:

$$L(\lambda) = P(y|X) \quad (7)$$

Where:

- N is the number of training examples.
- $X^{(i)}$ is the feature representation of sentence i.
- $y^{(i)}$ is the correct sentiment label sequence.

Training is done using Gradient Descent or L-BFGS Optimization.

Step 5: Inference (Predicting Sentiment Labels)

Given a new input sentence, the CRF predicts the most likely sequence of sentiment labels $\hat{y}$:

$$\hat{y} = arg\ max_y\ P(y|X) \quad (8)$$

Using Viterbi Decoding, the model finds the best sequence of sentiment labels.

Step 6: Compute Sentiment Counts

For a given text T_i, we count how many words belong to each category

$$P(T_i) = \left(\hat{y}_j = P\right) \quad (9)$$

$$N(T_i) = \left(\hat{y}_j = N\right) \quad (10)$$

$$Q(T_i) = \left(\hat{y}_j = Q\right) \quad (11)$$

Where:

- 1(.) is an indicator function that returns 1 if the condition is true, otherwise 0.
- P(T_i) is Positive Word Count
- N(T_i) is Negative Word Count
- Q(T_i) is Neutral Word Count

Advantages of CRF for Sentiment Analysis

1. Captures dependencies between words (unlike traditional classifiers).
2. Handles negation words (e.g., "not happy" is negative).
3. Works well for Aspect-Based Sentiment Analysis (ABSA).
4. Ensures sequence consistency (e.g., positive words likely follow each other).

3.4. HIERARCHICAL ATTENTION NETWORK (HAN)

Hierarchical Attention Networks (HAN) is a deep learning model designed to capture the hierarchical structure of text by applying attention at both the word level and sentence level. This allows the model to focus on the most important words within a sentence and the most relevant sentences within a document for sentiment classification.

ALGORITHM

Input:

- A dataset D containing n textual reviews {T_1, T_2,…, T_n}.
- Each review T_i consists of s sentences S={S_1, S_2,…, Sn}.
- Each sentence S_k consists of m words {w_1, w_2,)…, w_m}.

Output:

- A final sentiment classification y for each text T_i.

Step 1: Tokenization into Words & Sentences

Each document T_i is first split into sentences, and then each sentence is split into words:

$$T_i=\{T_1, T_2,\ldots, T_n\}, S_k=\{w_1, w_2,\ldots,w_m\} \quad (12)$$

Step 2: Word Embedding using Bi-GRU

Each word w_j is converted into a dense vector representation using a pretrained word embedding:

$$E(w_j) \in R^d \quad (13)$$

Where:

- d is the embedding dimension.
- $E(w_j)$ represents the semantic meaning of the word.

A Bidirectional Gated Recurrent Unit (Bi-GRU) is then applied to capture contextual dependencies:

$$h_j = B_i\text{-}GRU(E(w_j)) \quad (14)$$

Where:

- h_j is the hidden representation of word w_j.
- Bi-GRU allows long-term dependencies between words to be captured.

Step 3: Word-Level Attention Mechanism

Not all words contribute equally to sentiment. A word-level attention mechanism assigns importance scores to words:

$$u_j = tanh\left(W_w h_j + b_w\right) \tag{15}$$

$$\alpha_j = \frac{exp\left(u_j^T u_w\right)}{\sum_j exp\left(u_j^T u_w\right)} \tag{16}$$

$$S_k = \sum_j \alpha_j h_j \tag{17}$$

Where:

- W_w and b_w are trainable weight matrices.
- u_j is the word importance vector.
- α_j is the attention weight for each word.
- S_k is the sentence vector, a weighted sum of word representations.

This ensures that important words contribute more to the sentence representation.

Step 4: Sentence Encoding using Bi-GRU

Each sentence representation S_k is passed through another Bi-GRU to capture dependencies between sentences:

$$h_k = \text{Bi-GRU}(S_k) \tag{18}$$

Where:

- h_k is the hidden state representation of sentence S_k.

Step 5: Sentence-Level Attention Mechanism

Just like words, not all sentences are equally important. A sentence-level attention mechanism assigns weights to sentences:

$$u_k = tanh\left(W_s h_k + b_s\right) \tag{19}$$

$$\alpha_k = \frac{exp\left(u_k^T u_s\right)}{\sum_j exp\left(u_k^T u_s\right)} \tag{20}$$

$$V(T_i) = \sum_k \alpha_k h_k \tag{21}$$

Where:

- W_s and b_s are trainable weights.
- u_k is the sentence importance vector.
- α_k is the attention weight for each sentence.
- $V(T_i)$ is the final document representation.

This ensures that important sentences contribute more to the final sentiment classification.

Step 6: Classification Using Softmax

The final document representation $V(T_i)$ is passed through a fully connected layer with softmax activation to predict the sentiment:

$$y = Softmax(W_0 V(T_i) + b_o) \tag{22}$$

Where:

- W_s and b_s are trainable parameters.
- y is the predicted sentiment label.

Advantages of HAN for Sentiment Analysis

1. Captures Word & Sentence Hierarchy, Uses Bi-GRU to process words and sentences separately.
2. Attention Mechanism, learns which words and sentences are most important for sentiment.
3. Handles Long Documents, unlike models that only process words, HAN aggregates information across multiple sentences.
4. Better Interpretability, Attention scores provide insight into why the model made a prediction.

Fig. 1 illustrates the end-to-end pipeline of Sentivolve, which integrates FastText, Conditional Random Fields (CRF), and Hierarchical Attention Networks (HAN) for sentiment analysis.

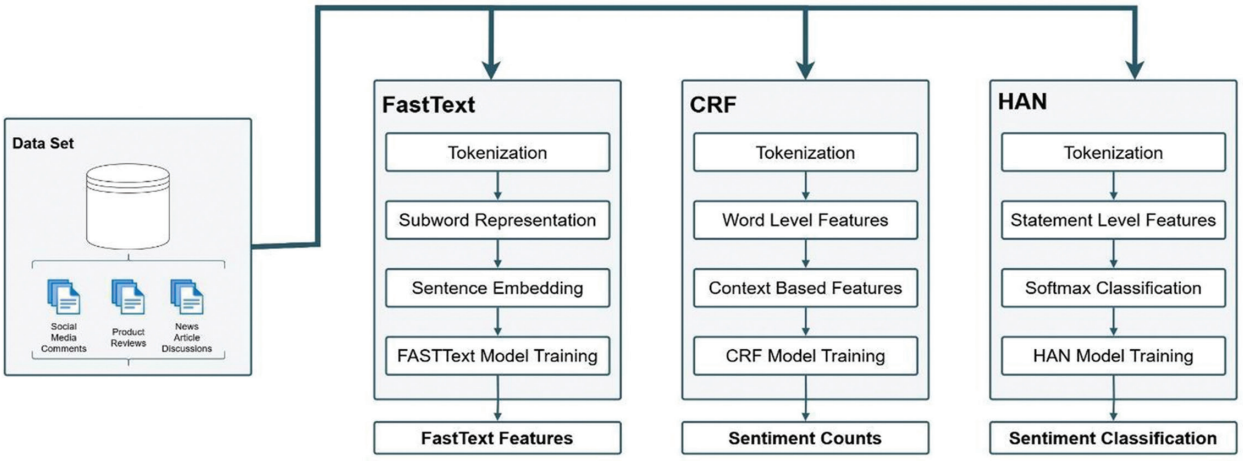


Fig. 1. Data flow diagram for FastText, CRF and HAN models

It begins with textual data from social media comments, product reviews, and news discussions, which undergo preprocessing and feature extraction through three distinct models. FastText generates word embeddings by capturing subword representations, CRF extracts word-level sentiment labels with contextual dependencies, and HAN identifies key sentence-level features using an attention mechanism.

These processed features contribute to sentiment classification, where FastText provides semantic understanding, CRF ensures word-level consistency, and HAN enhances interpretability. The combined features are then used in a Random Forest classifier, improving overall accuracy and robustness in sentiment prediction.

3.5. ENSEMBLE LEARNING WITH RANDOM FORESTS

Random Forest (RF) is an ensemble learning method that combines multiple Decision Trees to improve accuracy, reduce overfitting, and handle high-dimensional data. In sentiment analysis, RF is used to classify text into positive, negative, or neutral sentiments based on extracted features.

ALGORITHM:

Input:

- A dataset D containing n textual reviews $\{T_1, T_2, \ldots, T_n\}$.
- Each review T_i consists of s sentences $S=\{S_1, S_2, \ldots, S_n\}$.
- Each sentence S_k consists of m words $\{w_1, w_2, \ldots w_m\}$.
- Pretrained FastText model (word embeddings).
- CRF model trained for word-level sentiment labelling.
- HAN model trained for sentence-level sentiment representation.

Output:

- A final sentiment label y for each text T_i.

Step 1: Create Final Feature Vector

The final feature vector $X(T_i)$ combines FastText, CRF, and HAN features:

$$X(T_i) = \begin{bmatrix} V(T_i)_{FastText}, \\ P(T_i), N(T_i), Q(T_i), V(T_i)_{HAN} \end{bmatrix} \quad (23)$$

Where:

- $V(T_i)_{FastText}$ are sentence embeddings from FastText
- $P(T_i)$, $N(T_i)$, $Q(T_i)$ is Sentiment features from CRF
- $V(T_i)_{HAN}$ is attention-based sentence representation from HAN

Step 2: Train Random Forest with Bootstrapped Trees

- Each tree in Random Forest learns to classify sentiment using the feature vector:

$$y_k = f_k(X(T_i)) \quad (24)$$

Where:

- f_k is the decision function learned by tree k.
- y_k is the sentiment prediction by tree k.

Step 3: Final Sentiment Prediction using Majority Voting

The final sentiment label is obtained by majority voting across all trees:

$$\hat{y} = mode\{y_1, y_2, \ldots, y_K\} \quad (25)$$

Where:

- K is the total number of decision trees.
- $\hat{y}$ is the final sentiment classification.

Fig. 2 represents the Random Forest-based sentiment classification process, where data from social media comments, product reviews, and news discussions is first randomized to ensure diverse training samples.

The ensemble learning framework consists of multiple decision trees (Tree 1, Tree 2, …, Tree N), each independently trained on a different subset of the data. Each tree makes a sentiment prediction, and the final sentiment classification is determined through majority voting, ensuring improved accuracy and robustness against overfitting.

Fig. 3 illustrates how FastText, CRF, and HAN work together to classify textual data into Positive, Negative, or Neutral sentiment. The process begins with input text from sources such as social media comments, product reviews, and news discussions, which undergoes preprocessing before analysis.

First, CRF checks the positive word count; if it exceeds a predefined threshold, the sentiment is classified as Positive. If not, the HAN model evaluates sentence-level sentiment, and if the score is high enough, the text is still labeled as Positive. If both checks fail, CRF assesses negative word count, classifying the text as Negative if the count is above a certain threshold. If neither strong positive nor negative sentiment is detected, FastText computes semantic similarity, potentially shifting classification toward Positive.

If the similarity score is low, a final CRF check on neutral word count determines if the text should be classified as Neutral. This structured decision-making process leverages CRF for word-level sentiment tagging, HAN for sentence importance, and FastText for semantic understanding, ensuring a robust and context-aware sentiment classification system.

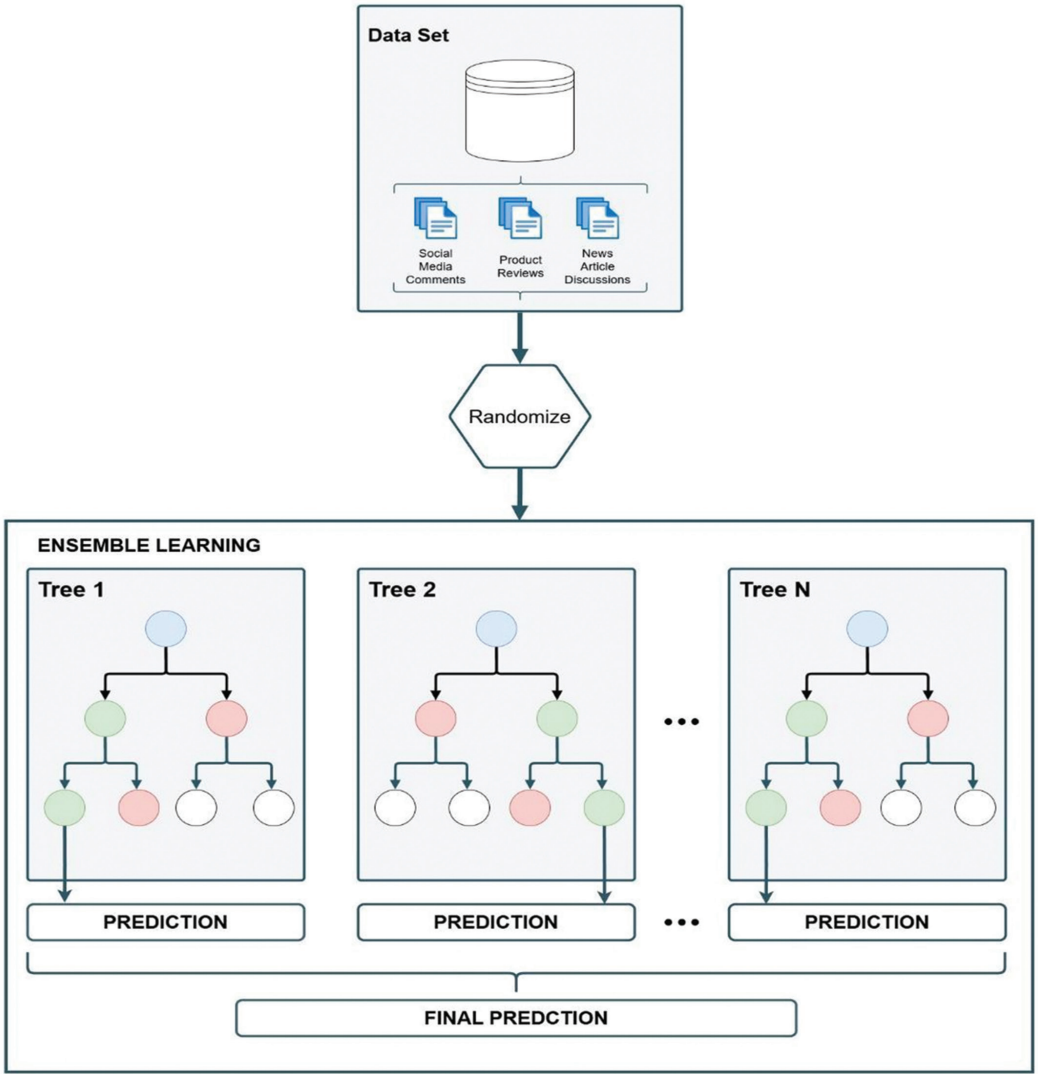


Fig. 2. Random Forest-based sentiment classification process

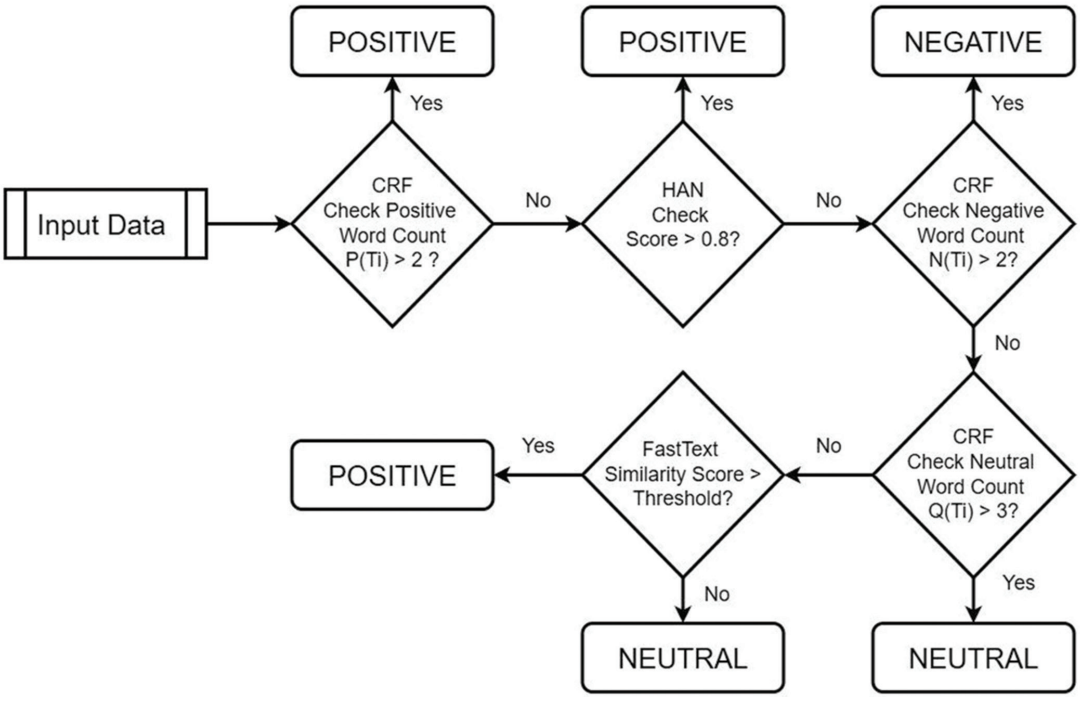


Fig. 3. Sentiment Classification Decision Process Using FastText, CRF, and HAN in RF

3.6. TRAINING STRATEGY AND COMPUTATIONAL CONSIDERATIONS

The components of the Sentivolve framework are trained independently, and their outputs are subsequently fused into a final feature representation used by the Random Forest classifier. Specifically:

1. The FastText model is either pretrained or trained separately to generate word embeddings.
2. The CRF model is trained using word-level features and sentiment labels.
3. The HAN model is trained to learn hierarchical representations at the word and sentence levels.
4. Finally, the output features from FastText, CRF, and HAN are concatenated and used as input to train the Random Forest classifier.

This modular training approach enables better component-wise tuning and flexibility in updating individual sub-models without retraining the entire pipeline. It also facilitates parallel processing during training, which can improve scalability across large datasets.

From a computational standpoint, integrating four distinct models does increase the overall training time and resource requirements compared to traditional or single-model pipelines. Specifically:

1. FastText training is relatively fast and memory-efficient due to its shallow architecture.
2. CRF training is moderately expensive, especially for large corpora with detailed word-level annotations.
3. HAN, being a deep learning model with Bi-GRU layers and attention mechanisms, is computationally intensive and benefits significantly from GPU acceleration.
4. The final Random Forest training step is less demanding but can scale in complexity with the size of the feature vectors produced by the previous modules.

On average, the combined training time for all components (including pre-processing and vector aggregation) was approximately 3.5 hours on a high-performance system with an NVIDIA RTX 3090 GPU and 32GB RAM. Despite this added cost, the performance gains—particularly in accuracy, recall, and F1-score—justify the computational investment.

4. RESULTS

Our experimental results demonstrate that integrating FastText, CRF, and HAN into a Random Forest model significantly enhances sentiment classification performance. The proposed hybrid approach outperforms traditional models by capturing semantic meaning, contextual dependencies, and sentence importance, leading to improved accuracy, precision, recall, and F1-score. This section presents the dataset selection, preprocessing steps, and evaluation metrics used to validate the effectiveness of our model.

4.1. DATASET SELECTION

To ensure a comprehensive evaluation, we employ publicly available sentiment analysis datasets from diverse domains, including:

- **Social Media Comments**: User-generated comments from platforms such as Twitter, Facebook, and Instagram, which provide informal and context-driven sentiment expressions.
- **Product Reviews**: A collection of user feedback on various products, categorized into positive, negative, and neutral sentiments.
- **News Article Discussions**: Sentiment-laden discussions and user comments on news articles, reflecting opinions on current events and trending topics.

These datasets provide a balance between short-form, informal texts (social media comments) and structured reviews and discussions (product reviews, news article discussions), allowing us to assess the model's generalizability across different text structures.

4.2. DATA PREPROCESSING

Before training our model, we apply the following preprocessing techniques to standardize the textual data:

- **Tokenization**: Splitting text into individual words or subwords.
- **Stopword Removal**: Eliminating common words (e.g., "the," "is," "and") that do not contribute to sentiment meaning.
- **Lemmatization**: Converting words to their base forms (e.g., "running" → "run").
- **Handling Negations**: Converting phrases like "not good" into a single token (e.g., "not_good") to retain sentiment context.
- **Text Vectorization**: Converting words into numerical representations using FastText embeddings.

4.3. FEATURE EXTRACTION

Our approach integrates three key techniques for feature extraction:

- **FastText**: Generates dense word embeddings to capture word semantics.
- **CRF**: Assigns word-level sentiment labels to identify positive, negative, and neutral terms.
- **HAN**: Applies attention mechanisms to emphasize the most relevant sentences in a document.

These extracted features are then combined into a structured representation before being passed into the Random Forest classifier for final sentiment classification.

4.4. EVALUATION METRICS

To assess model performance, we utilize standard classification metrics:

- **Accuracy**: Measures overall correctness of predictions.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} \quad (26)$$

- **Precision**: Evaluates the proportion of correctly predicted positive cases.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (27)$$

- **Recall**: Assesses how well the model captures all relevant instances.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (28)$$

- **F1-Score**: A harmonic mean of precision and recall, balancing both measures.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (29)$$

4.5. EXPERIMENTAL ENVIRONMENT

The experiments were conducted on a high-performance computing setup with the following configurations:

- **Processor**: Intel Core i7 / AMD Ryzen 9
- **RAM**: 16GB / 32GB
- **GPU**: NVIDIA RTX 3090 (for deep learning models)
- **Programming Language**: Python (TensorFlow, Scikit-Learn, NLTK, and FastText libraries)

4.6. CONTRIBUTION OF FASTTEXT, CRF, AND HAN TO DECISION TREES IN RANDOM FOREST

The inclusion of FastText embeddings improves accuracy by capturing word semantics. CRF enhances recall by identifying word-level sentiment polarity, while HAN improves precision by focusing on sentence-level importance.

Table 1 illustrates the effect of different feature sets on the performance of the Random Forest (RF) model for sentiment classification. When RF was used without FastText, CRF, or HAN, it achieved an accuracy of 82.1%, with moderate precision, recall, and F1-score values.

Incorporating FastText improved the performance significantly, increasing accuracy to 85.6%. The addition of CRF further boosted accuracy to 86.2%, while HAN contributed to a slightly higher performance at 87.1%. The best results were obtained when all three techniques—FastText, CRF, and HAN—were integrated with RF, achieving the highest accuracy of 90.4% and the best overall precision, recall, and F1-score values. These results indicate that combining multiple feature extraction and representation techniques enhance the model's ability to capture contextual and semantic information, leading to improved sentiment classification performance.

To provide a clearer understanding of these performance improvements, a graph has been created for visual representation in Fig. 4, making it easier to compare the impact of different feature sets on accuracy, precision, recall, and F1-score.

Table 1. evaluates the individual contributions of FastText, CRF, and HAN to sentiment classification.

Feature Set in RF	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
RF without FastText, CRF, HAN	8 2.1	80.9	81.4	81.1
RF with FastText	85.6	84.3	85.1	84.7
RF with CRF	86.2	85.0	85.7	85.3
RF with HAN	87.1	86.3	86.7	86.5
RF with FastText + CRF + HAN	90.4	89.8	90.2	90.1

4.7. PERFORMANCE COMPARISON OF SENTIVOLVE

Table 2 presents a comparative analysis of various sentiment classification models based on accuracy, precision, recall, and F1-score. Traditional models like Naïve Bayes and SVM achieved moderate performance, with accuracy values of 78.2% and 81.5%, respectively. Deep learning approaches such as BiLSTM and Transformer (BERT) significantly outperformed these traditional models, reaching 85.3% and 88.5% accuracy. Among all, the proposed hybrid model, which integrates Random Forest (RF) with FastText, Conditional Random Fields (CRF), and Hierarchical Attention Networks (HAN), exhibited the highest performance with an accuracy of 90.4%, along with superior precision, recall, and F1-score values.

Table 2. Compares the performance of traditional and deep learning models for sentiment classification

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Naïve Bayes	78.2	76.5	77.0	76.7
SVM	81.5	80.2	80.8	80.5
BiLSTM	85.3	84.1	84.9	84.5
Transformer (BERT)	88.5	87.7	88.2	88.0
Sentivolve	90.4	89.8	90.2	90.1

The results highlight the effectiveness of combining multiple techniques for robust sentiment classification, demonstrating that hybrid architectures can outperform both traditional machine learning and standalone deep learning models. A graph has been created for visual representation for the same in Fig. 5.

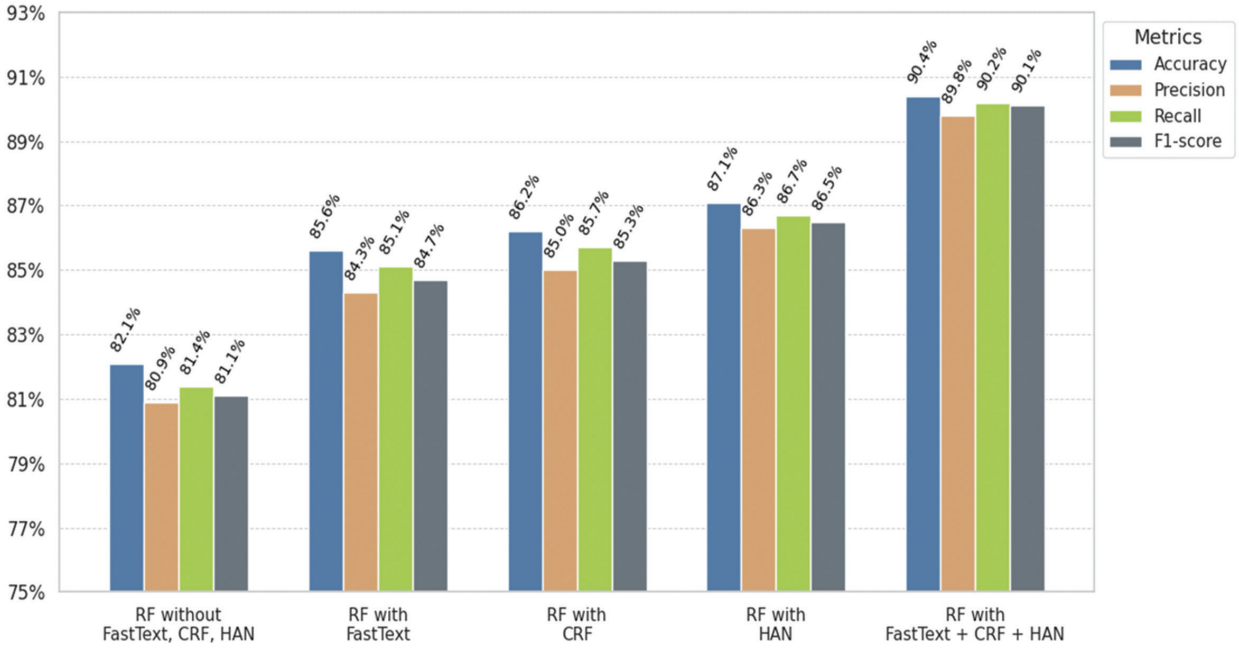


Fig. 4. graph comparison between the existing models

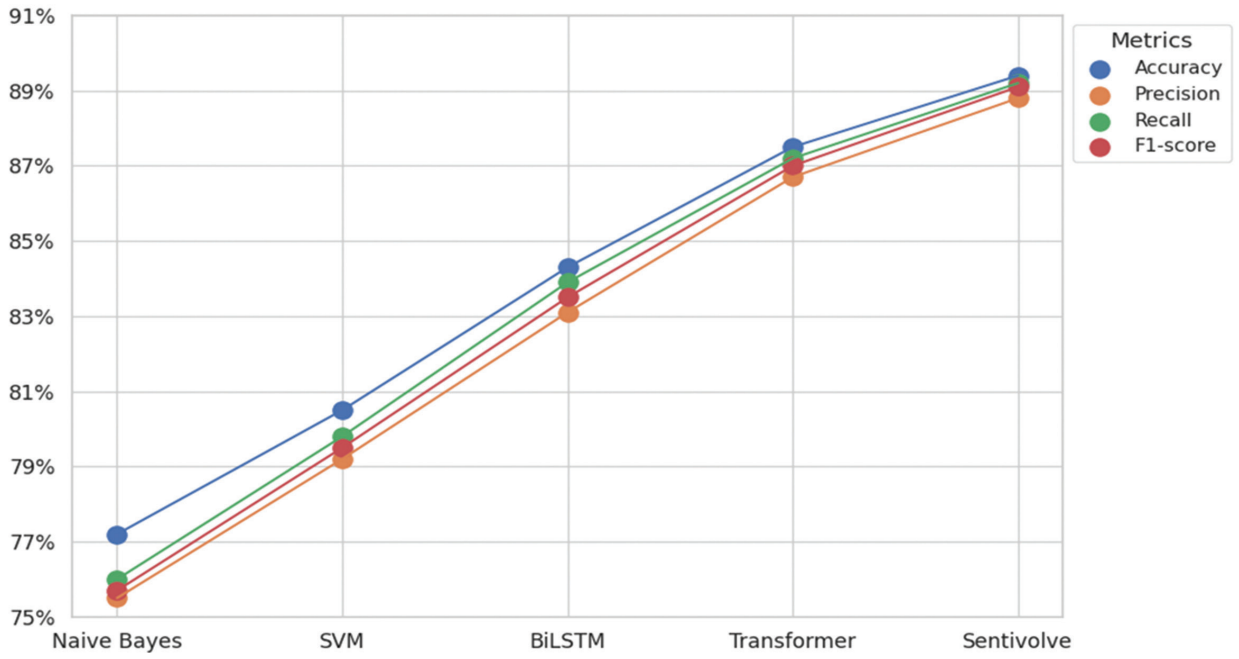


Fig 5. graph comparison between the existing models

4.8. ANALYSIS AND INTERPRETATION OF CONFUSION MATRICES

Tables 3, 4, and 5 present confusion matrices for sentiment classification across three different datasets: social media, customer reviews, and news articles. Each table compares the actual sentiment labels (Positive, Negative, Neutral) with the predicted classifications, highlighting the model's performance. In the social media dataset (Table 3), the model correctly classified 381 positive, 494 negative, and 484 neutral instances, while misclassifications were observed primarily between similar sentiment categories, such as 56 negative instances misclassified as positive and 31 neutral instances misclassified as negative. Similarly, for the customer review dataset (Table 4), the model demonstrated high accuracy, correctly predicting 410 positive, 470 negative, and 468 neutral sentiments, with minimal misclassifications. The

news article dataset (Table 5) followed a similar pattern, with 390, 485, and 470 correct classifications for positive, negative, and neutral sentiments, respectively. Across all datasets, the model maintained strong performance in distinguishing sentiment classes, although minor confusion between neutral and other classes suggests potential areas for improvement in handling ambiguous text.

To enhance interpretability, heatmap visualizations Figs. 6, 7, and 8 have been generated for each dataset, providing a graphical representation of misclassifications.

These heatmaps illustrate the distribution of correct and incorrect predictions, making it easier to identify patterns in the model's performance. While the overall classification accuracy remains high, the heatmaps highlight areas where the model occasionally confuses neutral sentiments with positive or negative classes, indicating potential improvements in handling ambiguous text.

The results demonstrate that combining multiple features significantly enhances sentiment classification accuracy. FastText provides rich word representations, CRF ensures word-level sentiment consistency, and HAN captures sentence importance, making Random Forest's decision process more robust. The hybrid model effectively balances contextual understanding, syntactic structure, and hierarchical text representation, leading to superior sentiment prediction.

4.9. CROSS-VALIDATION EVALUATION

To ensure the robustness and generalizability of the Sentivolve model, we employed 5-fold cross-validation as part of the performance evaluation. The dataset was randomly divided into five equal subsets. In each iteration, four subsets were used for training while the remaining subset was used for validation. This process was repeated five times, and the reported results represent the average performance across all folds as shown in Table 6.

Table 3. Social Media Dataset Confusion Matrix

Predicted / Actual	Positive	Negative	Neutral
Positive	**381**	56	19
Negative	45	**494**	19
Neutral	23	31	**484**

True Positives: 381 | True Negatives: 494 | True Neutral: 484

Table 4: Customer Review Dataset Confusion Matrix

Predicted / Actual	Positive	Negative	Neutral
Positive	**410**	32	22
Negative	36	**470**	18
Neutral	25	30	**468**

True Positives: 410 | True Negatives: 470 | True Neutral: 468

Table 5: News Article Dataset Confusion Matrix

Predicted / Actual	Positive	Negative	Neutral
Positive	**390**	48	18
Negative	39	**485**	22
Neutral	20	29	**470**

True Positives: 390 | True Negatives: 485 | True Neutral: 470

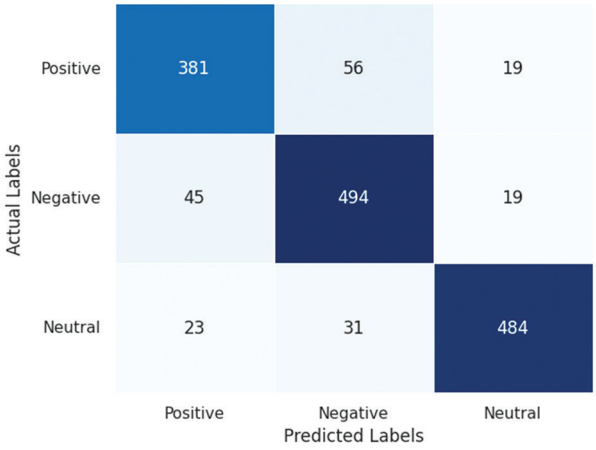


Fig. 6. Social Media Dataset Heatmap

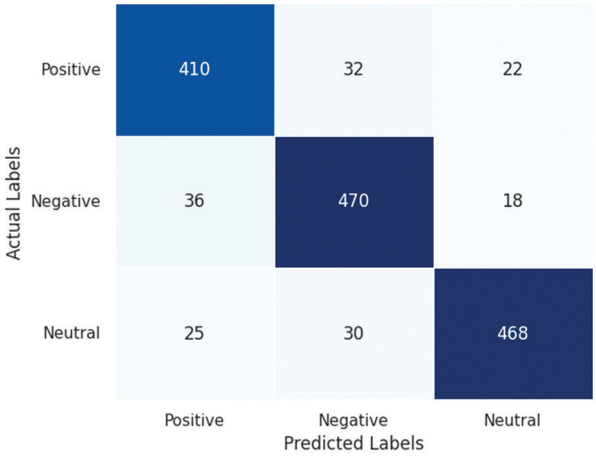


Fig. 7. Customer Review Dataset Heatmap

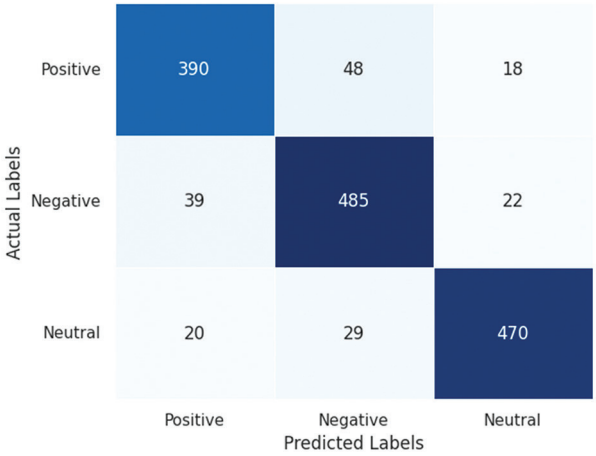


Fig. 8. News Article Dataset Heatmap

Table 6. Dataset Confusion Matrix

Fold	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	90.1	89.6	89.9	89.7
2	90.4	89.8	90.2	90.1
3	90.7	90.1	90.5	90.3
4	90.3	89.9	90.1	90.0
5	90.5	90.0	90.3	90.2
Average	90.4	89.8	90.2	90.1

These results reaffirm that Sentivolve consistently achieves high sentiment classification performance across different data splits, underlining its reliability and suitability for real-world deployment.

Our hybrid sentiment classification model, leveraging FastText, CRF, HAN, and Random Forest, achieves state-of-the-art accuracy. Future work will focus on integrating transformer-based models such as BERT with our existing approach to further improve interpretability and classification performance. Additionally, optimizing computational efficiency while maintaining high accuracy will be explored for large-scale real-time applications.

5. CONCLUSION

The Sentivolve system demonstrates significant advancements in sentiment analysis by integrating multiple advanced techniques—FastText embeddings, Conditional Random Fields (CRF), Hierarchical Attention Networks (HAN), and Random Forests. The fusion of these diverse methods enables the system to achieve superior performance in terms of accuracy, precision, recall, and F1-score, confirming its robustness and adaptability across different textual domains.

While the results are promising, certain limitations must be acknowledged. First, the modular architecture, though flexible, introduces added complexity in model integration and feature fusion. Each component requires separate training and fine-tuning, which can increase development overhead and model management effort. Second, the computational cost and training time are significantly higher compared to lightweight, single-model approaches—especially due to the HAN and CRF components, which demand GPU acceleration and sequence-level processing. Additionally, the reliance on handcrafted feature combinations for Random Forest classification may limit scalability when transitioning to real-time applications or multilingual settings.

Another limitation is that the current model architecture does not fully leverage transformer-based embeddings like RoBERTa or DeBERTa, which have shown superior context modeling in recent literature. Finally, while Sentivolve performs well on benchmark datasets, its generalizability to noisy or low-resource languages has not yet been validated.

Future research will focus on addressing these limitations by:

1. Exploring end-to-end training pipelines to simplify integration,
2. Incorporating transformer-based components for improved contextual understanding,
3. Optimizing model runtime for real-time inference, and Extending evaluation to multilingual and domain-specific datasets.

Despite these challenges, Sentivolve offers a robust foundation for sentiment analysis tasks, especially in scenarios requiring high precision and interpretability. Its modular design, while complex, allows targeted improvements and paves the way for scalable, hybrid sentiment analysis systems.

6. REFERENCES

[1] M. Idris, A. Rifai, K. D. Tania, "Sentiment Analysis of Tokopedia App Reviews using Machine Learning and Word Embeddings", Sinkron: Jurnal dan Penelitian Teknik Informatika, Vol. 9, No. 1, 2025.

[2] H. Naceri, N. Hicham, K. Satori, "Enhancing Arabic E-commerce Review Sentiment Analysis using a Hybrid Deep Learning Model and FastText Word Embedding", EAI Endorsed Transactions on Internet of Things, Vol. 10, 2024.

[3] L. Yao, N. Zheng, "Sentiment Analysis Based on Improved Transformer Model and Conditional Random Fields", IEEE Access, Vol. 12, 2024, pp. 90145-90157.

[4] A. Yadav, V. Singh, "DLL5143A@NLU of Devanagari Script Languages 2025: Detection of Hate Speech and Targets using Hierarchical Attention Network", Proceedings of the First Workshop on Challenges in Processing South Asian Languages, Abu Dhabi, UAE, 2025, pp. 278-288.

[5] A. Yadav, S. S. Rathore, "A Hierarchical Attention Networks based Model for Bug Report Prioritization", Proceedings of the 17th Innovations in Software Engineering Conference, Bangalore, India, 22-24 February 2024, pp. 1-5.

[6] P. Putrayasa, E. Utami, R. Marco, "Comparing Algorithms in Sentiment Analysis on DUKCAPIL App Reviews on Playstore Using Ensemble Learning Methods", G-Tech: Jurnal Teknologi Terapan, Vol. 9, No. 1, 2025, pp. 190-201.

[7] G. Yadav, M. U. Bokhari, S. I. Alzahrani, S. Alam, M. Shuaib, "Emotion-Aware Ensemble Learning (EAEL): Revolutionizing Mental Health Diagnosis of Corporate Professionals via Intelligent Integration of Multi-modal Data Sources and Ensemble Techniques", IEEE Access, Vol. 13, 2025, pp. 11494-11516.

[8] T. Alsaedi, M. R. R. Rana, A. Nawaz, A. Raza, A. Alahmadi, "Sentiment Mining in E-Commerce: The Transformer-based Deep Learning Model", International Journal of Electrical and Computer Engineering Systems, Vol. 15, No. 8, 2024, pp. 641-650.

[9] M. Rahman, Y. Watanobe, M. A. Islam, "RoBERTa BiLSTM: A Context Aware Hybrid Model for Sentiment Analysis", Journal of Big Data, Vol. 9, 2024, p. 12.

[10] A. Jahin, S. H. Shovon, M. F. Mridha, M. R. Islam, Y. Watanobe, "TRABSA: A Hybrid Transformer Attention BiLSTM for Robust and Interpretable Sentiment Analysis of Tweets", Expert Systems with Applications, Vol. 208, 2024, p. 118224.

[11] M. M. Hossain, M. S. Hossain, M. F. Mridha, M. Safran, S. Alfarhood, "Multi Task Opinion Enhanced Hybrid BERT Model (Opinion BERT) for Sentiment and Mental Health Classification", Scientific Reports, Vol. 15, 2025, p. 3332.

[12] F. Ullah, S. Faizullah, I. U. Khan, T. Alghamdi, T. A. Syed, A. B. Alkhodre, M. S. Ayub, A. Karim, "Prompt Based Fine Tuning with Multilingual Transformers for Language Independent Sentiment Analysis", Scientific Reports, Vol. 15, 2025, p. 20834.

[13] F. Alqarni, A. Sagheer, A. Alabbad, H. Hamdoun, "Emotion Aware RoBERTa Enhanced with Emotion Specific Attention and TF IDF Gating for Fine Grained Emotion Recognition", Scientific Reports, Vol. 15, 2025, p. 17617.

[14] M. Rahman, A. Islam, Y. Watanobe, E. Zekaoui, "Analysis of the Evolution of Advanced Transformer-Based Models: Experiments on Opinion Mining", Opinion Mining & Sentiment Analysis, Vol. 45, 2024, pp. 77-95.

[15] T. Zekaoui, S. Yousfi, M. Rhanoui, M. Mikram, "Transformer-Based Opinion Mining: A Comparative Study", Artificial Intelligence Review, Vol. 58, 2024, pp. 1335-1354.

[16] A. K. Ullah, N. Rana, Z. Wang, "ECO SAM: Emotion Correlation Enhanced Sentiment Analysis Model", Frontiers in Psychology, Vol. 15, 2024, p. 210189.

[17] M. S. Islam, R. Kumar, A. Sharma, "Challenges and Future in Deep Learning for Sentiment Analysis: A Comprehensive Review and Novel Hybrid Approach", Artificial Intelligence Review, Vol. 57, No. 3, 2024, pp. 2301-2335.

[18] P. Zarin, L. Chen, "Enhancing Twitter Sentiment Analysis Using Hybrid Transformer BiLSTM Architectures", Expert Systems, Vol. 41, No. 4, 2024, p. e12.

[19] T. Nguyen, M.-V. Truong, K. Nguyen, "Fine Grained Cross Modal Fusion Framework for Sentiment Analysis", Information Fusion, Vol. 80, 2024, pp. 120-135.

[20] A. Tran, P. Smith, J. Lee, "Emotion Aware Hybrid BERT for Mental Health and Sentiment Classification", IEEE Transactions on Affective Computing, Vol. 16, 2025, pp. 55-68.

[21] R. Kaseb, M. Farouk, "SAIDS: Sentiment Analysis Informed of Dialect and Sarcasm", Arabian Journal for Science and Engineering, Vol. 48, No. 1, 2023, pp. 341-358.

[22] M. Mir, B. Zhang, C. Liu, "Hierarchical Attention Networks with BERT Embeddings for Sentiment Analysis", IEEE Access, Vol. 11, 2023, pp. 23456-23467.

[23] S. Ullah, A. Faisal, S. Khan, "Multilingual Sentiment and Emotion Detection Using XLM R and mBERT", Journal of Computational Science, Vol. 58, 2024, p. 101348.

[24] R. Ullah, S. Khan, M. Kim, "Hybrid Transformer CNN BiLSTM Models for Enhanced Twitter Sentiment Analysis", Journal of Information Processing Systems, Vol. 20, No. 1, 2024, pp. 45-60.

[25] F. Ullah, S. Faizullah, I. U. Khan, "Prompt Based Sentiment Analysis for Low-Resource Languages", International Journal of Computer Applications, Vol. 182, No. 20, 2023, pp. 1-9.

[26] A. S. Rahman, M. U. Khatun, "Transfer Learning for Sentiment Classification Using BERT in Multilingual Contexts", Sensors, Vol. 23, No. 11, 2023, p. 5232.

[27] M. Yadav, M. Vishwakarma, "Hybrid Ensemble Learning for Multi Lingual Sentiment Classification", Scientific Reports, Vol. 14, 2024, p. 11234.

[28] J. Davis, R. Patel, "Tensor-Based Fusion BERT for Multimodal Sentiment Analysis", Neurocomputing, Vol. 500, 2025, pp. 120-134.

A Video Summarization Technique using Multi-Feature DWHT and GMM for CBVR System

Original Scientific Paper

Dappu Asha*
Jawaharlal Nehru Technological University Hyderabad,
Department of Electronics and Communication Engineering
Telangana, India
ashamanickrao@gmail.com

Y. Madhavee Latha
Malla Reddy Engineering College for Women, affiliated to JNT University,
Department of Electronics and Communication Engineering
Telangana, India
madhaveelatha2009@gmail.com

*Corresponding author

Abstract *– The increasing utilization of multimedia data and digital information in present times presents a vast scope for research in content-based retrieval systems. An improved CBVR System is proposed to extract video streams effectively using DWHT Multi-features and GMM. Our CVBR method performs VSBD for identifying Video shots by computing DWHT on video frames for multi-feature extraction, and then key frames are identified. A summarized frame is developed using the VS algorithm based on GMM on the UCF Dataset. Later, a procedure is applied for the input query video stream, and correlation coefficients are calculated between the query and the database multi-feature vectors, giving us similarity measures. Lastly, our experimental results validate the efficiency of our proposed CBVR System, achieving an average precision of 0.821 and a loss of 0.179, outperforming existing CBVR systems using DCT and optimized perceptual VS, which have precision values of 0.6475 and 0.71, respectively, along with losses of 0.3525 and 0.29.*

Keywords*: Content-based Video Retrieval (CBVR), Discrete Walsh-Hadamard Transform (DWHT), Video Shot Boundary Detection (VSBD), Video Summarization (VS), Gaussian Mixture Model (GMM).*

Received: June 30, 2025; Received in revised form: September 19, 2025; Accepted: September 22, 2025

1. INTRODUCTION

Advancements and improvements in technology have made a large amount of information available on the web [1]. Due to this, the demand for automatic tools for browsing, retrieving, intelligent surveillance, and ranking of information has gained importance [2]. Since video is a significant source of information available on the web, it occupies a large memory size and requires machinery for analyzing [3]. Content-based retrieval is essential because text-based retrieval is limited by human errors and manipulations [4, 5]. The first two levels in the CBVR framework are Video Shot Boundary Detection (VSBD) and Video Summarization (VS). A video shot is an assembly of similar frames formed by still or moving camera images [6]. The detection of the transition from one shot to the next shot is called shot detection. The shot transitions are categorised into CT (Cut Transition) and GT (Gradual Transition) [7]. CT is an abrupt change between one video shot and the succeeding video shot, whereas GT is a slow change that occurs in the video stream, and it continues for many video frames that arises due to video editing. Several kinds of video editing effects exist, such as dissolve, fadeout, fade-in, etc [7]. The method of mechanically segmenting a video stream into video shots or scenes is termed VSBD [7]. VS is a crucial step in the CBVR system, reducing the video's dimensionality to a single frame.

Our proposed CBVR System comprises online and offline processes. The offline process is carried out on database videos, and the online process is computed on the query video. There are four steps in the CBVR process. It starts with VSBD to identify shots and key frames, then VS is applied to summarize the key frames. Next, from the summarized frame, DWHT-based multi-

features are extracted, and lastly, similarity is measured to retrieve similar videos.

The framework of our paper is ordered as follows: Section 2 bounces on a literature review of the current CBVR techniques. In Section 3, our proposed CBVR system is illustrated. In Section 4, experimental results are presented. Finally, Section 5 discusses the conclusion of our work.

2. LITERATURE REVIEW

Numerous Automatic Video Retrieval systems have been proposed in the past few years. From the literature reviews, different retrieval methods like text-based [8], content-based, query image-based [9], and sketch-based [10] were developed. The Literature survey is presented in Table 1, providing a brief overview of features, datasets, results, advantages, and limitations.

Table 1. Literature Survey on CBVR System

Author /Year of publication	Search Type	Features	Methods/ Techniques	Database and Results	Advantages	Limitations
Palanivelu et al. (2024) [11]	Query input	Pertinent visual features using ResNet50	CNN	TREC02, TREC10, YTAD09, and IDV01 Accuracy of 58.33, 91.67, 92.08, and 23.08, respectively	CNN can automatically learn complex features from raw video data	Poor performance on certain types of complex datasets, the computational cost is high and requires more labelled data
Farhan et al. (2021) [12]	Query by Example	Color features	Discrete Cosine Transform (DCT)	Real World 8 Classes each contain videos Precision of 0.6475	Effective and automatic feature extraction from video content	Did not consider semantic features and evaluated on a small database
Sathiyaprasad et al. (2020) [13]	Query input	I-GLCM (Improved Gray Level Co-Occurrence Matrix)	RPCNN (Region-based Pre-Convoluted Neural Network)	MNIST (4000 images), KAGGLE Precision of 0.9067	Combines I-GLCM and R-PCNN, which aims to optimize accuracy	Using local identifiers and descriptors increases the computational cost
Dyana et al. (2010) [14]	Query by Example	MST-CSS (Multi-Spectro-Temporal Curvature Scale Space)	Multiscale and multispectral Filters	480 real-world video Shots 50 classes each contain 20 videos Precision of 0.71	Combining shape contour and motion trajectory through multiscale and multispectral processing	Operates only on static backgrounds
Shivanand et al. (2019) [15]	Query Input	Semantics contents	ROI and ACF Detector	Own 70 video Dataset captured from a mobile phone No evaluation metric considered	Focuses on techniques for detecting the Region of Interest (ROI)	Self-collected dataset, ROI detection is primarily focused on the signboards only
Mallick et al. (2019) [16]	Query Input	Motion Vector	Spatial Pyramid matching (Haar Transformation -4 Level)	UCF Dataset VCD Dataset Precision of 0.8862	Ultilizes motion vector-based key frame extraction as a video summarization technique to recapitulate video content	Fails in guaranteeing its efficiency for conspicuous motion
Thomas et al. (2019) [17]	Query Input	Single frame-based approach	Human visual system, optimization	Standard video data sets UCF, MED, CCV, BBC, OVP Precision of 0.71	Uses a single summarized frame for indexing instead of multi-frame indexing, the method reduces computational complexity and memory demands for video databases	Incapability lies in its sensitivity to segmentation, background extraction errors, and its inability to effectively summarize crowded foreground activities
Asha et al. (2018) [18]	Query Input	Color Distributions Texture & Motion Binary Patterns	LBP and SAD	40 videos from Google 4 classes each contain 10 videos Precision of 0.80	A multiple-feature approach improves performance	High Computational cost leading to more execution time

Reference [11] highlights the significant advancements in Content-Based Video Retrieval through the application of deep learning, particularly using models like Inception ResNet. It enables the extraction of intricate, high-level features from video frames, leading to more precise and efficient video search and retrieval capabilities. Reference [12] used a transform-based CBVR system grounded on single DCT color features; the study achieved an average result of 0.6475 by implementing DCT on a database comprising 100 videos across various categories, with 5 videos in each category. These findings underscore the efficacy of using DCT in video retrieval processes. while [13] illustrated region-based Improved-GLCM on RPCNN using image

query, this approach aims to address the computational constraints and accuracy limitations of existing CBVR systems. Reference [14] discusses a unified approach using MST-CSS attributes generated by multi-spectral filters to efficiently represent video objects. This approach integrates spatial and temporal information within a unified framework. Reference [15] explains the semantic features and ROI, which require additional user inputs. Reference [16] is based on motion vector features and spatial pyramid matching; it narrows the search space. This approach is motivated by its ability to overcome limitations of Bag-of-Features methods by considering feature spatial layout. It involves partitioning an image or key frame into finer sub-regions and computing local features for each sub-region. [17] explains video summarization using human perception and optimization for redundancy removal. It highlights the growing need for efficient video summarization due to increased video consumption. It identifies shortcomings in existing methods, particularly their inability to accurately represent video events and adapt to different scene types. The proposed solution focuses on a context-driven, perceptually optimized framework that creates a single summarized frame, promising enhanced retrieval performance and reduced resource demands. [18] signifies the use of multiple features for improving retrieval, it highlights that an effective CBVR system requires considering both spatial and temporal features to achieve accurate results, distinguishing it from Content-Based Image Retrieval (CBIR).

3. IMPLEMENTATION OF PROPOSED CBVR SYSTEM

In this section, we enlighten on our CBVR system. VSBD is the first step in the CBVR system where DWHT is applied. The second step is Video Summarization (VS) using a Gaussian Mixture Model (GMM), and the third step is video retrieval using the extracted Multi-Features. The block diagram of the proposed CBVR System is shown in Fig.1.

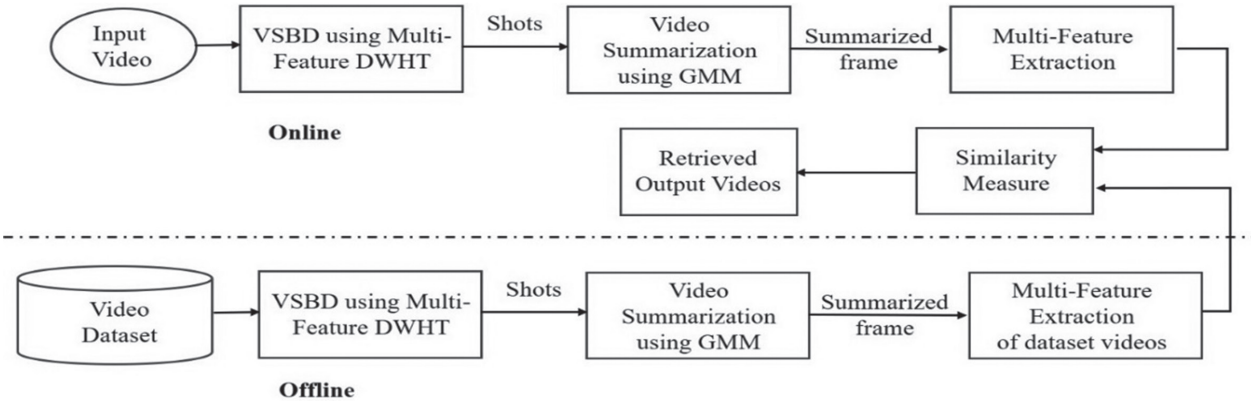


Fig. 1. Proposed CBVR System Block Diagram

3.1. WALSH-HADAMARD TRANSFORM

Discrete Walsh Hadamard Transform (DWHT) [19] is immensely used in numerous applications of image and video processing because of its robustness, energy compaction, fast computation, less memory storage space, and flexibility. DWHT is defined below:

Let $f_i(x, y)$ be the i^{th} frame of size MxN. The forward discrete Walsh-Hadamard $X_i(u, v)$ can be expressed as in (1)

$$X(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f_i(x,y)g(x,y,u,v)$$
$$for\ x,y\ are\ spatial\ coordinates\ and\ u,v\ are\ coordinates\ in\ transform\ domain \quad (1)$$
$$x,u = [0,1,2\ldots.M-1] and\ y,v = [0,1,2\ldots.N-1]$$
$$g(x,y,u,v) is\ fordward\ Mask$$

The forward kernel of DWHT is defined as in (2)

$$g(x,y,u,v) = \frac{1}{N}(-1)^{\sum_{i=0}^{m-1}[b_i(x)p_i(u)+b_i(y)p_i(v)]} \quad (2)$$

Where N=2m is the size of the transform matrix. The summation in exponent is performed in modulo 2 arithmetic, and $b_i(y)$ is the ith bit in the binary representation of y.

The DWHT matrix of order 8 (N=8) is shown in (3)

$$For\ N=8: \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \quad (3)$$

The blending functions of DWHT are characterized in mask vectors as $W = \{w_1, w_2, ..., w_{64}\}$ aligned from top to bottom and left to right of the DWHT masks for N=8, given in Fig. 2. These masks of the DWHT help in extracting multi-feature vectors.

3.2. VSBD TECHNIQUE

The VSBD process is accomplished in four steps: computing DWHT kernels, multi-feature extraction, composing a continuous vector, and identifying the video shot boundary. Various VSBD techniques used previously are presented here. Reference [20] projected a technique for extracting key frames of multi-features, the algorithm leverages deep prior information and

multi-feature fusion to enhance saliency extraction. [7] Proposed content-based VSBD using the Haar transform to accurately detect abrupt and gradual video transitions. [21] combines candidate segment selection with SVD for dimensionality reduction and employs distinct pattern matching techniques. [22] discusses WHT and a procedure-based identification process to distinguish all shot transitions, and [23] discusses cut detection using a histogram where a single feature is considered and the GT is neglected. We observe that transform-based techniques are more accurate, and multi-feature extraction with a procedure-based approach will help in detecting CT and GT. Video is read into the system, and multi-features are extracted by protrusive DWHT kernels on each video frame. The dissimilarity and similarity among succeeding video frames are measured by calculating the correlation between consecutive feature vectors. The procedure-based VSBD algorithm is useful to identify shot transitions. The block diagram of the VSBD Technique is shown in Fig. 3.

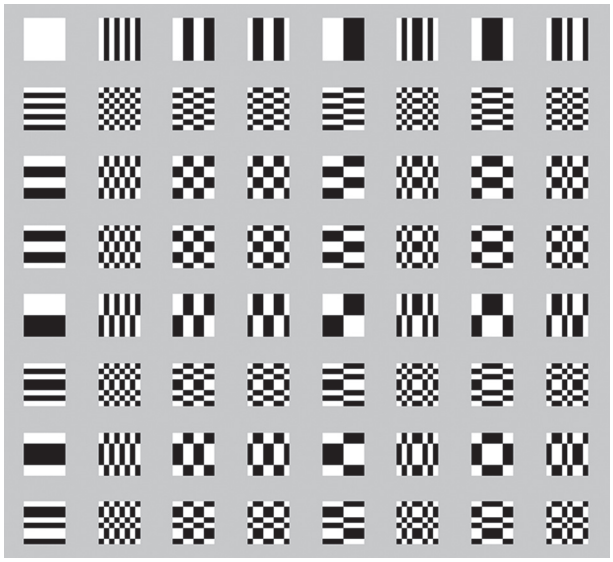


Fig. 2. DWHT (Discrete Walsh-Hadamard Transform) Kernels

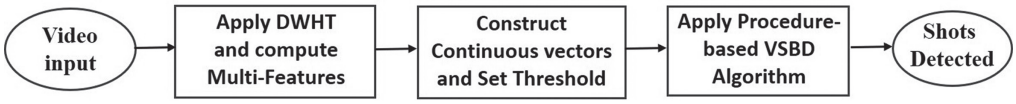


Fig. 3. Block diagram of the VSBD Technique

3.3. MULTI-FEATURE VECTOR EXTRACTION

In the video shot boundary detection, the multi-feature vector extraction phase is very significant. We extract features like motion, color, shape, and texture vectors by applying DWHT blending functions on video streams.

The blending functions or kernels used for color, shape, and texture feature extraction are shown in equation (4). We use kernel w_2 to w_{64} for high-frequency and w1 for low-frequency demonstrations. The kernels w_1, w_{33}, and w_{37} are shown below in equation (4):

$$W_1 = w_1 = \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$W_2 = w_{33} = \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix} \quad (4)$$

$$W_3 = w_{37} = \frac{1}{8}\begin{bmatrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

The shape feature vectors are computed by projecting the w_{33} and w_{37} masks. The texture feature vectors are computed by projecting w_1, w_{33}, and w_{37} masks.

The motion feature vectors are calculated by following the steps

1. Project W_4=DWHT kernel for N=8 as in (3) on succeeding frames.
2. Calculate Motion Vector (MV) using the SAD (Sum of Absolute Difference) method.
3. Extract motion features by subtracting MV from the projected W_4 successive frames.
4. Calculate the correlation between succeeding motion strength frames.

3.4. FEATURE EXTRACTION PROCEDURE

Let $X_m = \{x_{m1}, x_{m2}, x_{m3}, x_{m4}\}$, X_m as in (5), signify the anticipated values of blocks by designing the inner product of K (K_m) where m = [1,2, … .*No. of blocks*] and W_j, j=1, 2, 3, 4, respectively.

X_m are computed using below equation:

$$X_m = \{x_1^m = \langle K_m, W_1\rangle,\ x_2^m = \langle K_m, W_2\rangle, x_3^m = \langle K_m, W_3\rangle, x_4^m = \langle K_m, W_4\rangle\} \quad (5)$$

Where $\langle K_m, W_1 \rangle = \sum_{i=1}^{p} K_{mi} * W_{ji}$.

Here, W_{ji} is the i^{th} value of W_j blending vector, K_{mi} is the i^{th} value of K_m, and p is the number of pixels in each block K_m.

i. The Color Feature Vector (C_m) of the consequent block is obtained as in (6):

$$C_m = x_1^m = \langle K_m, W_1\rangle \quad (6)$$

ii. The Shape Feature Vector (E_m) of the consequent block is obtained as in (7):

$$E_m = \sqrt{(x_2^m)^2 + (x_3^m)^2} \quad (7)$$

iii. The Texture Feature Vector (T_m) of the consequent block is obtained as in (8) and (9):

$$T_m = |K_m^2 - Z^2|$$
$$Z = x_1^m W_1 + x_2^m W_2 + x_3^m W_3 \quad (8)$$

$$K_m = \sum_{i=1}^{3} \langle K_m, W_i \rangle W_i \quad (9)$$

iv. The Motion Feature Vector (M_m) of the consequent block is obtained as in (10):

$$M_m = |x_4^m - ME| \quad (10)$$

3.5. CONSTRUCTION OF CONTINUOUS VECTOR

Subsequently, after identifying multi-features, the next phase in the VSBD involves computing a continuous vector to determine the similarity between successive frames, as outlined in the equations below, where P represents the Correlation coefficient. The estimated correlation coefficients between the succeeding frames are calculated among the blocks of the fi and fi+1 frames [21]. Hence, for all feature vectors of color (C), Texture (T), Shape (S), and Motion (M), the equivalent continuous vector equations are given as in (6-10) [23].

After constructing individual features, a continuous vector as in (11-14), we calculate the mean of all individual vectors as in (15). Continuous vector is in the range of [0,1]. On these combined coefficients (μ), a procedure-based VSBD algorithm is applied for recognizing shot transitions.

$$\alpha(f) = P(f, f+1) = \sum_{m=1}^{no\ of\ blocks} corrcoef\left(C_{m,f} - C_{m,f+1}\right) \quad (11)$$

$$\beta(f) = P_S(f, f+1) = \sum_{m=1}^{no\ of\ blocks} corrcoef\left(E_{m,f} - E_{m,f+1}\right) \quad (12)$$

$$\gamma(f) = P_T(f, f+1) = \sum_{m=1}^{no\ of\ blocks} corrcoef\left(T_{m,f} - T_{m,f+1}\right) \quad (13)$$

$$\delta(f) = P(f, f+1) = \sum_{m=1}^{no\ of\ blocks} corrcoef\left(M_{m,f} - M_{m,f+1}\right) \quad (14)$$

$$\mu(f) = \frac{1}{4}\{\alpha(f) + \beta(f) + \gamma(f) + \delta(f)\} \quad (15)$$

3.6. PROCEDURE FOR THE VSBD ALGORITHM

Our proposed procedure for the VSBD algorithm is grounded on the following guidelines to identify the CT and GT. If the sequential frames are identical, then the continuous vector will be high and when frames differ, the values will be low. A Threshold value (T_h) is computed to identify the shots, and T_h is calculated by taking the mean of the continuous signal as in (16).

$$Th = \frac{1}{n}\left(\sum_{f=1}^{n} \mu(f)\right) \ where\ n = No.of\ Frames\ in\ video. \quad (16)$$

According to the continuous correlation coefficient values, it is easy to recognize the presence of CT. The continuous correlation values for GT in the video sequence between the successive frames will be lower. An example of a VSBD plot is shown in Fig. 4.

Procedure for finding Valley points:

1. Compute all continuous signal $\mu(f)$ as in (15) where f=1,2…..n, n=No. of frames.

   Calculate the Threshold value (Th) by using equation (16).

   Find the valley points $V(k)$ that are less than the threshold Th.

2. Location of the valley is stored as $Lop(m) = k$, where the valley occurred at the k^{th} frame, $m = m+1$, till finding all valley points, and the procedure ends.

Procedure-based Shot Detection

Step 1: Read $\mu(f)$, $V(k)$, nv: number of valleys, Lop: location of valley point.

Step 2: Set the Threshold (Th).

Step 3: If the values before and after the valley point are greater than Th, then identify the shot transition as CT.

Step 4: Else identify the shot as GT.

Step 5: If the valley point values are in a gradual transition and are less than 0.6, then identify as a fade transition.

Step 6: Compare the valley point values with the previous values and decrement $c2$ until the condition is false, where $c2$ represents the starting point of the fade transition.

Step 7: Compare the valley point values with the upcoming values, incrementing $c1$ until the condition is false. Here, $c1$ represents the end point of the fade transition.

Step 8: If the valley point values in GT exceed 0.6, then identify it as a dissolve transition.

Step 9: Compare the valley point values with the previous values and decrement $c2$ until the condition is false, where $c2$ represents the starting point of the dissolve transition.

Step 10: Compare the valley point values with the upcoming values, incrementing $c1$ until the condition is false. This represents the end point of the dissolve transition.

Step 11: Repeat from step 3.

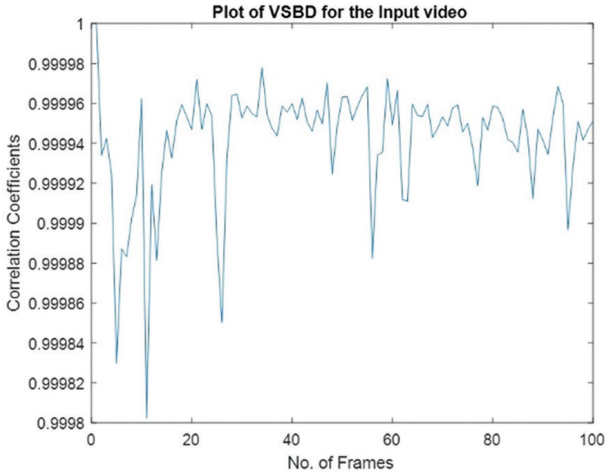


Fig. 4. VSBD plot

3.7. VS USING GMM

VS using the GMM method for detecting foreground [24, 25] is illustrated here. GMM is built for foreground Modelling. It is an extensively common procedure for moving object detection. It executes a soft clustering method to categorize each pixel as foreground or background by assigning a score to each pixel indicating the strength of the pixel [26, 27].

GMM is computed on each pixel using equations (17) and (18) below.

$$P(X_t) = \sum_{i=1}^{K} w_{i,t} \cdot \Omega(X_t, \mu_{i,t}, \sigma_{i,t}) \tag{17}$$

where Xt : pixel in t^{th} frame

K: the number of components

$w_{i,t}$: weight of the K^{th} component in t^{th} frame

$\mu_{i,t}$: the mean of K^{th} component in t^{th} frame

$\sigma_{i,t}$: the standard deviation of K^{th} component in t^{th} frame

where $\Omega(X_t, \mu_{i,t}, \sigma_{i,t}$ probability density function

$$\Omega(X_t, \mu, \sigma) = \frac{1}{(2\pi)^{1/2} |\sigma|^{1/2}} \, exp^{-\frac{1}{2}(X_t - \mu)\sigma^{-1}(X_t - \mu)} \tag{18}$$

The Procedure for VS using the GMM Algorithm is given below, and the flowchart is shown in Fig.5.

1. Read frames from video.
2. Extract shots using VSBD using Multi-Feature DWHT.
3. Extract key frames from shots, apply foreground detection using the GMM method, and create a mask.
4. Apply 2D Gaussian and morphological filtering for smoothing and noise removal (the outputs of different stages are shown in Fig.6).
5. Select the frame and stitch objects on it to get the summarized frame shown in Fig.7.

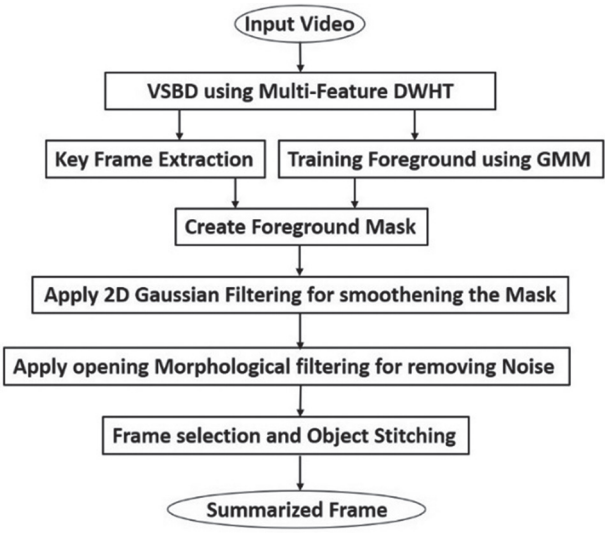


Fig. 5. Flowchart for VS Technique using GMM

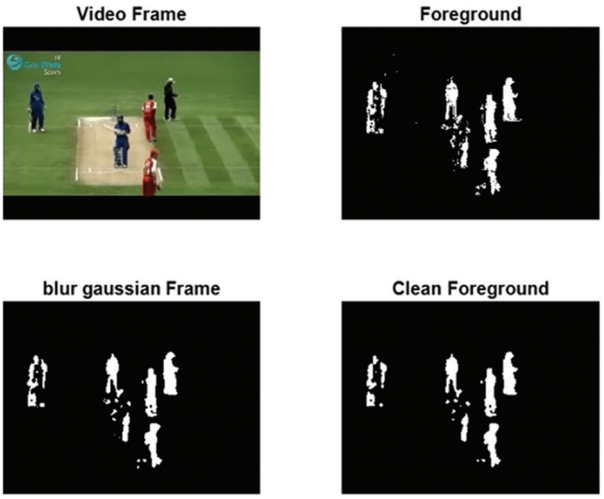


Fig. 6. VS using GMM Algorithm Outputs

Fig. 7. Summarized Frame output

3.8. MULTI-FEATURE EXTRACTION

The proposed VSBD method splits the video stream into scenes or shots, and we select the key frames from each shot, considering the middle frame of a shot as a key frame. Apply the VS Algorithm to get a summarized frame. The multi-features of the summarized frame are extracted using equations (5-10), and a feature vector for that video stream is formed. This procedure of extracting a feature vector for all videos in the UCF da-

tabase [28] is done in offline mode. In online mode, this procedure is applied to the query video. Next, we perform a similarity measure by calculating correlation coefficients. The top 10 videos with high correlation coefficients are retrieved and displayed.

4. EXPERIMENT RESULTS

The performance of the CBVR System is tested on the UCF database [28], consisting of human action videos. We considered 20 classes, each with 10 videos, totalling 200 videos. Table 2 shows the properties of the database. A few examples of retrieved videos for the given query are shown in Table 3. The performance is evaluated using Precision (Pr), Loss, Compression Ratio (CR) [29], and online Execution Time (ET). The precision, loss, and CR are intended to be used with the equations (19-21). The superior precision values enhance the performance of the CBVR system. The average precision and loss for the UCF dataset are shown in Fig. 8. Table 4 discusses the comparison of our proposed CBVR system with other systems.

$$Precision\ (P_r) = \frac{Correct\ no.of\ videos\ retrieved}{Total\ no.of\ videos\ retrieved} \quad (19)$$

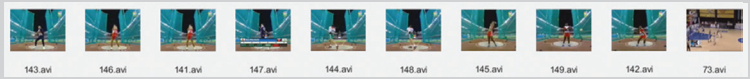
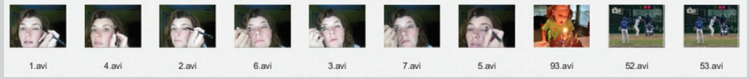
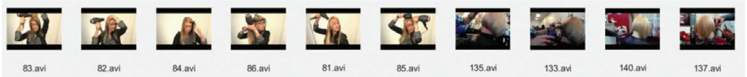
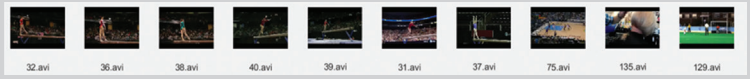
$$Loss\ = \frac{Incorrect\ no.of\ videos\ retrieved}{Total\ no.of\ videos\ retrieved} \quad (20)$$

$$Compression\ Ratio\ (CR) = 1 - \left(\frac{No.of\ Key\ frames}{Total\ no.of\ frames}\right) \quad (21)$$

Table 2. UCF Video Database Properties and Description

Description	Quantity
No. of Videos in Database	200
Average Duration	8.7513 seconds
Frame Rate	30 fps
Resolution	240x320
Average No. of Frames	153
Video Format	avi
Bits per pixel	24
Total Implementation time to extract features	1154.79832 seconds

Table. 3. Examples of Retrieved Video streams for the given query

Class	Query	Retrieved videos	Pr	Loss	CR	ET (Sec)
Cricket Shot			0.9	0.1	0.89	29.83
Field Hockey Penalty			0.8	0.2	0.95	29.24
Hammer Throw			0.9	0.1	0.92	21.18
Baseball Pitch			0.7	0.3	0.95	20.55
Apply Eye Makeup			0.7	0.3	0.92	20.27
Blow Dry Hair			0.6	0.4	0.94	23.97
Balance Beam			0.7	0.3	0.88	21.09

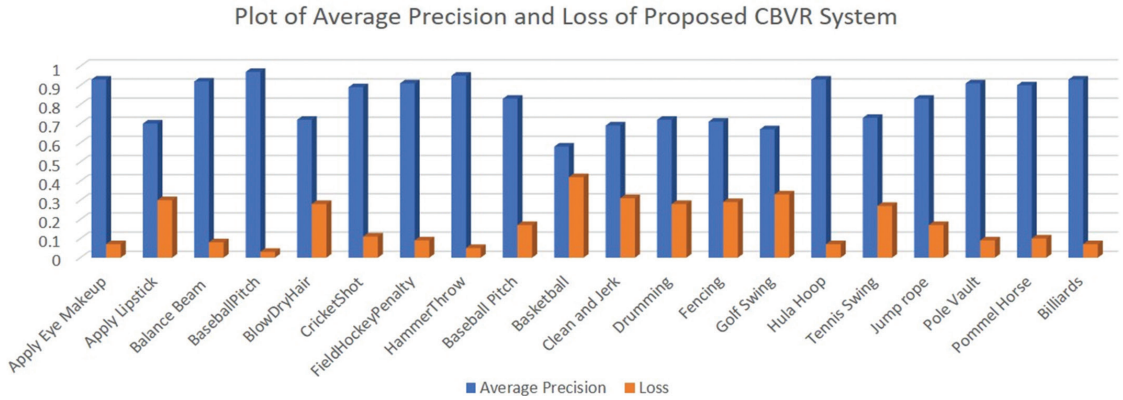


Fig. 8. Plot of Average Precision and Loss of different Video Classes

Table 4. Comparison of the other CBVR Systems to our Proposed work

Methods	Average Precision	Loss
CBVR using DCT [12]	0.6475	0.3525
CBVR using optimized perceptual video Summarization [17]	0.71	0.29
Proposed CBVR System	0.821	0.179

5. CONCLUSION

This paper proposes a novel method for CBVR, utilizing Multi-Feature DWHT and VS with the GMM Algorithm. From a video sequence, we first calculate the DWHT Multi-feature vector, and the correlation between successive frames is plotted. A procedure-based VSBD algorithm is used to divide the video into shots. Secondly, key frames are extracted, and the foreground is detected from them. A summarized frame is then stitched using GMM. From the summarized frame, multi-features are extracted and correlation coefficients between query and dataset videos are computed to retrieve similar videos. Experiments are performed on the UCF dataset, and the proposed CBVR system is evaluated. The proposed CBVR system has an average precision of 0.821 and a loss of 0.179, showing the performance of our work. In the future, we can improve the performance by making the system robust to camera motions and illumination variations.

6. REFERENCES

[1] A. Moutaoukkil, A. Idarrou, I. Belahyane, "Information retrieval approaches: A comparative study", International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 10, 2022, pp. 961-970.

[2] W. V. Ramos, A. P. Pumaleque, J. G. Torres, "Bibliometric Analysis of Scientific Production of Intelligent Video Surveillance", International Journal of Electrical and Computer Engineering Systems, Vol. 16, No. 6, 2025, pp. 461-471.

[3] A. S. Adly, I. Hegazy, T. Elarif, M. S. Abdelwahab, "Development of an Effective Bootleg Videos Retrieval System as a Part of Content-Based Video Search Engine", International Journal of Computing, Vol. 21, No. 2, 2022, pp. 214-227.

[4] G. S. N. Kumar, V. S. K. Reddy, L. K. Balivada, "Content-Based Video Retrieval Using Deep Learning Algorithms", Intelligent Systems and Sustainable Computing, Vol. 363, Springer, 2023, pp. 557-568.

[5] W. Hu, N. Xie, L. Li, X. Zeng, S. Maybank, "A survey on Visual Content Based Video Indexing and Retrieval", IEEE Transactions, On System, Man, And Cybernetics-Part C: Applications and reviews, Vol. 41, No. 6, 2011, pp. 797-819.

[6] A. Hussain, M. Ahmad, T. Hussain, I. Ullah, "Efficient content-based video retrieval system by applying AlexNet on key frames", ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, Vol. 11, No. 2, 2022, pp. 207-235.

[7] D. Asha, Y. M. Latha, "Content-Based Video Shot Boundary Detection Using Multiple Haar Transform Features", Advances in Intelligent Systems and Computing, Vol. 900, Springer, 2019, pp. 703-713.

[8] A. S. Adly, M. S. Abdelwahab, I. Hegazy, T. Elarif, "Issues and Challenges for Content-Based Video Search Engines A Survey", Proceedings of the 21st International Arab Conference on Information Technology, Giza, Egypt, 28-30 November 2020, pp. 1-18.

[9] A. Araujo, B. Girod, "Large-Scale Video Retrieval Using Image Queries", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 28, No. 6, 2018, pp. 1406-1420.

[10] L. Wang, X. Qian, X. Zhang, X. Hou, "Sketch-Based Image Retrieval With Multi-Clustering Re-Ranking", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 30, No. 12, 2020, pp. 4929-4943.

[11] P. Nitish, J. Nishita, D. Nishith, J. Bharati, "Content Based Video Retrieval using Deep Learning", Research Square, 2024, pp. 1-20.

[12] S. Hamad, A. S. Farhan, D. Y. Khudhur, "Content based video retrieval using discrete cosine transform", Indonesian Journal of Electrical Engineering and Computer Science, Vol. 21, No. 2, 2021, pp. 839-845.

[13] B. Sathiyaprasad, K. Seetharaman, B. S. Kumar, "Content based Video Retrieval using Improved Gray Level Co-Occurrence Matrix with Region-based Pre-Convoluted Neural Network–RPCNN", Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, Thoothukudi, India, 3-5 December 2020, pp. 558-563.

[14] A. Dyana, S. Das, "MST-CSS (Multi-Spectro-Temporal Curvature Scale Space) a Novel Spatio-Temporal Representation for Content-Based Video Retrieval", IEEE Transactions on Circuits and Sys-

tems for Video Technology, Vol. 20, No. 8, 2010, pp. 1080-1094.

[15] S. S. Gornale, A. K. Babaleshwar, P. L. Yannawar, "Analysis and Detection of Content based Video Retrieval", International Journal of Image, Graphics and Signal Processing, Vol. 11, No. 3, 2019, pp. 43-57.

[16] A. K. Mallick, S. Mukhopadhyay, "Video Retrieval Based on Motion Vector Key Frame Extraction and Spatial Pyramid Matching", Proceedings of the 6th International Conference on Signal Processing and Integrated Networks, Noida, India, 7-8 March 2019, pp. 687-692.

[17] S. S. Thomas, S. Gupta, V. K. Subramanian, "Context Driven Optimized Perceptual Video Summarization and Retrieval", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 29, No. 10, 2019, pp. 3132-3145.

[18] D. Asha, Y. M. Latha, V. S. K. Reddy, "Content Based Video Retrieval system using Multiple Features", International Journal of Pure and Applied Mathematics, Vol. 118, No. 14, 2018, pp. 287-294.

[19] R. C. Gonzalez, R. E. Woods, "Digital Image Processing", Second edition, Pearson, 2019.

[20] Q. Zhong, Y. Zhang, J. Zhang, K. Shi, Y. Yu, C. Liu, "Key Frame Extraction Algorithm of Motion Video Based on Priori", IEEE Access, Vol. 8, 2020, pp. 174424-174436.

[21] Z.-M. Lu, Y. Shi, "Fast Video Shot Boundary Detection Based on SVD and Pattern Matching", IEEE Transactions on Image Processing, Vol. 22, No. 12, 2013, pp. 5136-5145.

[22] G. L. Priya, S. Domnic, "Walsh-Hadamard Transform Kernel-Based Feature Vector for Shot Boundary Detection", IEEE Transactions on Image Processing, Vol. 23, No. 12, 2014, pp. 5187-5197.

[23] G. L. Priya, S. Domnic, "Video Cut Detection using block-based Histogram Differences in RGB Color Space", Proceedings of the International Conference on Signal and Image Processing, Chennai, India, 15-17 December 2010, pp. 29-33.

[24] F. Joy, V. Vijayakumar, "An improved Gaussian Mixture Model with post-processing for multiple object detection in surveillance video analytics", International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 8, 2022, pp. 653-660.

[25] A. Lajari, R. Sachin, "Dealing Background Issues in Object Detection using GMM: A Survey", International Journal of Computer Applications. Vol. 150, No. 5, 2016, pp. 50-55.

[26] A. Nurhadiyatna, W. Jatmiko, B. Hardjono, A. Wibisono, I. Sina, P. Mursanto, "Background Subtraction Using Gaussian Mixture Model Enhanced by Hole Filling Algorithm (GMMHF)", Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Manchester, UK, 13-16 October 2013, pp. 4006-4011.

[27] K. M. Angelo, "A novel approach on object detection and tracking using adaptive background subtraction method", Proceedings of the Second International Conference on Computing Methodologies and Communication, Erode, India, 15-16 February 2018, pp. 1055-1059.

[28] K. Soomro, A. R. Zamir, M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild", CRCV-TR-12-01, November, 2012.

[29] D Rajeshwari, V. Priscilla C, "An Enhanced Spatio-Temporal Human Detected Keyframe Extraction", International Journal of Electrical and Computer Engineering Systems, Vol. 14, No. 9, 2023, pp. 985-992.

Privacy Integrity-Aware Blockchain Communication in Federated Edge Learning Platform

Original Scientific Paper

Chitresha Jain*
Pandit Deendayal Energy University,
Computer Science & engineering Department
Raysan, Gandhinagar, India
Chitresha.research@gmail.com

Payal Chaudhari
Pandit Deendayal Energy University,
Computer Science & engineering Department
Raysan, Gandhinagar, India
Payal.Chaudhari@sot.pdpu.ac.in

*Corresponding author

Abstract *– The blockchain architecture offers transparent security mechanisms in a decentralized manner; due to this, it has attained increasing growth in a federated edge-server learning environment. In federated learning, the data model is executed in multiple edge servers in a collaborative manner, increasing users' privacy and data-integrity breach because of single point failure attack in the main computational server. Blockchain employing a rewarding mechanism in federated edge-learning platform aids the model to overcome single-point aggregation failure. However, the current method failed to identify selfish and baized workers; further, reaching global consensus model to assure privacy-integrity in blockchain-enabled federated edge-server is difficult. This paper presents privacy-integrity-aware blockchain communication (PIABC) in federated edge-server learning platform. The PIABC model is very effective in comparison with existing blockchain-privacy preserving schemes for identifying the correctly aggregated packets and eliminating malicious packets within the federated edge-server learning platform.*

Keywords*: Blockchain, Consensus, Federated Learning, Integrity, Secure aggregation, Privacy*

Received: May 8, 2025; Received in revised form: August 28, 2025; Accepted: August 29, 2025

1. INTRODUCTION

In the past few years, there has been a notable increase in the popularity of Federated-Learning (FL) [1]. Artificial-Intelligence (AI) like Machine-Learning (ML) and Deep Learning (DL) approaches, can be trained immediately on devices used by users as well as at edge of network using FL, which eliminates the need to centralize unprocessed information [2]. As a result, data breaches are less likely to occur, and users' privacy is protected whenever confidential data is stored on their devices. Additionally, when employees collaborate, they can access a wealth of information, which enhances efficiency and makes FL models more adaptable and effective. However, despite these advantages, FL also presents several challenges and limitations [3]. The primary features of FL make it vulnerable to novel attacks, which include (i) system-heterogeneity; (ii) the necessity of a reliable centralized entity to coordinate analysis of locally-trained approaches; (iii) vulnerable to inference attacks and information counterfeiting; (iv) absence of a reward approach for involved nodes; (v) communication-security; along with (vi) regulating issues [3, 4]. Moreover, scholars have started looking into methods that facilitate the utilization of blockchain since the FL method's present implementation lacks the necessary capabilities to deal with such issues [3, 5]. Both the public and private sectors are interested in FL because of its endless possibilities, which arise because of decentralized framework. FL depends on the likelihood of carrying out transactions that are legitimate and verifiable without requiring the participation of an unauthorized third-party, while also assuring the tracking and storage of information securely. Therefore, the integration of blockchain along with FL enhances the existing framework, guaranteeing the protection of private information, reliability, and framework safety in decentralized collaborating-learning applications [6].

Moreover, while the FL and Edge-Computing (EC) environment provides data-privacy and data-security preserving frameworks, it still faces challenges and threats, which are mentioned below [7, 8]. ***Data security attacks***: as the FL is executed in internet-of-things and multi-edge server in a collaborative manner there is a higher chance of data being attacked thus impacting data integrity and various security vulnerability by giving access to unauthorized person. Hence, one of the most important things to think about when designing FL security approaches is how to make a trusted framework in a place that is unreliable. ***Data Privacy-Preserving Problems***: Since the task is executed across different service nodes on the edge, there is a higher chance of privacy leaks. Thus, researchers studying privacy-preserving FL approaches across EC face new challenges due to the ever-changing nature of attack types. ***Computation and Communication Overhead***: The swift proliferation of Smart-IoT (SIoT) services has resulted in a significant increase in volume of information at edge-nodes, resulting in increased computation and communication cost [9]. Thus, exploration of new FL approaches in EC environments is constrained by the inadequate computing efficiency, restricted transmission bandwidth, real-time networking, and high standards of service demands of edge-devices. ***Diverse attack***: the FL is prone to different attack like Free-Rider and Poisoning Attacks, Sybil, and inference attack [10]. Thus, there is a need for a more enhanced model that deals with different kinds of security attacks. ***Single-Point Failures***: FL exhibits vulnerability towards single-point failures due to its reliance on a central server for the transmission of model variables required for updating the model. FL frameworks, when integrated with blockchain technologies, enhance local decentralization and provide an efficient approach. To keep information stored securely, blockchain nodes work together. Their combined abilities allow them to check all stored models and information for malicious activity on any node [9]. However, the current blockchain model poses certain challenges in reputation design in detecting poisoning and backdoor attack behavior [10].

To address the above issue, several studies have been published in the literature [11-18] that utilize particular reward processes. The fundamental concept of the current reward or incentive-based mechanisms is that individuals provide modified information by introducing noise to maintain integrity and privacy, while fusion-centers compensate for the compromise of users' integrity and privacy [11, 12]. Nonetheless, this brings forth two additional challenges:

(i)determining an appropriate level of noise to maintain the necessary privacy and (ii) ensuring effective information trustworthiness while ultimately reducing the effects of compromised information. It is essential to initially measure the threshold to preserve integrity and privacy, followed by concurrent improvement of aggregation accuracy while ensuring that users are provided with suitable thresholds for integrity and privacy preservation [12]. However, creating a reward system that guarantees integrity is essential. The compensation of users who provide altered confidential information about their respective preserved privacy-integrity threshold established by the federated coordinator is essential [16]. This indicates that the compensation coming from aggregation/fusion-centers to users is connected with the user's trust, which is derived from the dependability of their submitted information through an effective verification procedure [17, 18]. This work introduces an effective learning-driven approach for privacy-integrity aware blockchain communication (PIABC) for a federated edge-learning platform, aimed at addressing the aforementioned challenges. The proposed model offers the best possible trade-off among fusion accuracy and integrity-privacy preservation level, resulting in optimum integrity-privacy preservation data fusion. Then, it validates the dependability of the information and finally updates worker and user trust (reputations) and calculates the fused weights accordingly. **The contributions of work are as follows**:

- The paper introduces an innovative blockchain-based communication model designed to ensure both data privacy and integrity in federated edge learning.
- A novel trust mechanism is developed, utilizing blockchain for secure user authentication while preserving user privacy.
- A global consensus model is designed to rigorously enforce privacy and integrity standards within the federated edge learning framework.
- The model demonstrates higher throughput, improved detection rates, and fewer misclassifications of attacks compared to existing methods.

The paper is organized as follows: Section 2 discusses various current methods designed to provide blockchain-enabled federated learning to mitigate different attacks. The section highlighted the benefits and limitations of the current security mechanism for a federated edge learning platform. Section 3 aims at designing a novel approach to address both privacy and integrity issues, adopting a trust and consensus model. Section 4 validates the result of the proposed approach over existing methodologies. Finally, the significance of research in terms of performance parameters is discussed alongside the future direction to enhance the security model.

2. LITERATURE SURVEY

This section reviews security schemes for federated learning (FL) ensuring user privacy and data integrity, emphasizing their contributions and limitations. H. Liu *et al.* [11] proposed a blockchain-based trust model using a trust-computing sandbox and state-channel blockchain with smart contracts to handle malicious activity and task scheduling via Deep Reinforcement Learning (DRL). Simulations with the OPENAI GYM framework showed

improvements in task completion, cost, and SLA, but lacked security evaluation. W. E. Mbonu *et al.* [12] introduced a blockchain-enabled secure aggregation method to protect central servers, improve scalability, and reduce single points of failure, while using fault-tolerant servers for stragglers and callbacks to cut training time and storage. Evaluations on MNIST confirmed better accuracy, lower communication cost, and scalability, but the model did not fully address attacks.M. A. Mohammed *et al.* [13] presented an approach for healthcare, where they utilized blockchain-based FL for scheduling and offloading data to central servers, presenting an energy-efficient model called Energy-Efficient Distributed FL Offloading-Scheduling (ED- FOS). The main aim was to reduce energy, training time, and provide better Quality-of-Service. Simulations showed that EDFOS minimized energy consumption by 39%, training duration by 29% and resource consumption by 36%. This EDFOS provided better outcomes for training FL, yet failed to provide any security for users. M. Zirui *et al.* [14] presented a blockchain-based privacy-preserving approach for the healthcare sector, i.e., to help users overcome depression because of COVID-19. This work utilized a consensus blockchain-based privacy-preserving approach for providing security, privacy, trust, and interoperability. For simulation, a blockchain environment was created and evaluations were conducted in terms of cost, latency, and trust, where the best outcomes in comparison with existing approaches were achieved, yet failed to provide any outcomes on providing security. H. Javed *et al.* [15] presented a security model for monitoring systems used in smart healthcare. Their main aim was to handle insider malicious attacks. Hence, this work focused on providing security in the cloud for presenting a model called Cloud-Access Security-Broker (CASB), which collected all actions (logs) performed by users and provided security using blockchain. Evaluations were conducted by simulating an environment where patients' data was collected, and whenever a user retrieved the data, the user id was evaluated. Results were evaluated in terms of data storage duration and overall blockchain performance. The proposed approach provided integrity, scalability, privacy, accessibility, and transparency. S. T. Ahmed *et al.* [8] presented an approach for smart healthcare that utilized blockchain-based FL. Their main aim was to provide privacy by using a global-based aggregation approach, which indexed data in a central server and synchronized the knowledge server. Evaluations were conducted by considering various IoT devices, where the evaluated IoT behavior and classification delay were considered. Each IoT device was labeled and delay was evaluated within the FL environment, and findings show that their approach reduced labeling time and delay.

I. U. Din *et al.* [16] presented a trust-based approach called Context-Aware-Cognitive Memory-Trust Management-System (CACMTM) for smart transportation. The trust interactions among vehicles were constructed using game theory, where different trust modules were built. The main focus of this work was to utilize past trust established with vehicles (IoT nodes) to understand the behavior of vehicles, hence improving the trust approach. Also, considered a historical trust module to reduce attacks. After the trust module, it utilized blockchain for providing better security, accountability, and transparency. The CACMTM architecture was built in the following manner: first, trust was evaluated between vehicles, then trust was decided (i.e., to establish a connection or not), then update trust modules (all different trust modules), and finally utilize the trust in blockchain to provide security and prevent attacks. Simulations were conducted using the popular simulator OMNet++, where different attacks were evaluated. Evaluations were conducted in terms of accuracy, computation overhead, attack detection, and time, where CACMTM achieved the best results when compared with other approaches. Also, they evaluated results for different trust thresholds and for different numbers of IoT nodes, where the best result was obtained. Q. Xie *et al.* [17] presented an approach for the Internet of Vehicles (IoV) to provide security and identify attacks. This work utilized a cryptography key encryption and decryption approach along with Physical-Unclonable-Functions (PUFs) to identify an attack. Evaluations were conducted in terms of cost and how they can handle different attacks. Findings show that the approach can identify inside and outside attacks efficiently. Z. Ma *et al.* [18] presented an approach for IoV that was reliant on Road-Side-Units (RSUs), for which they presented a blockchain-based security-distributed authentication approach. This approach first collected data, preprocessed and stored data at the edge to decrease delay in communication and response time, which was done using a trusted authority. Further, smart contracts were used for authentication along with an enhanced Practical-Byzantine Fault-Tolerant Consensus approach for providing authentication for the blockchain ledger. Further, provided security using a Real-or-Random approach. Performance was measured in terms of execution and communication cost. However, the current method failed to identify selfish and biased workers; further, reaching a global consensus model to assure privacy-integrity in blockchain-enabled federated edge-server is difficult. In the next section, addressing the research core issues, the following methodology is presented.

3. PROPOSED METHODOLOGY

In this study, we propose a novel, blockchain-enabled, trust-based privacy-preserving authentication model to enhance security and privacy in federated learning (FL) environments. By leveraging blockchain technology, the system ensures secure and verifiable interactions between users and nodes. The model integrates dynamic trust evaluation mechanisms that continuously assess and update trust levels of participating nodes, based on recent interaction data and connection metrics. Blockchain's immutable ledger ensures transparency and accountability in this process, enabling real-time adaptation to node behavior and effectively isolating malicious entities.

Simultaneously, the methodology includes a privacy and integrity-aware consensus-based aggregation scheme to safeguard data privacy and integrity during model training. Blockchain technology underpins the aggregation process, providing a transparent and secure environment for model fusion. By incorporating secure computation techniques and differential privacy, the system ensures that the aggregation models remain robust against adversarial attacks and privacy breaches. The fusion of blockchain with federated learning provides a secure, trustworthy, and privacy-preserving framework, boosting the overall resilience and reliability of the system.

3.1. ARCHITECTURE

The architecture of the Privacy-Integrity-Aware Blockchain Communication (PIABC) framework, applied within a federated edge-server learning platform, is illustrated in Figure 1. This architecture leverages blockchain technology to establish a secure federated learning environment. The system consists of user nodes and computing nodes interconnected through the blockchain. Initially, users interact with the system via a blockchain-based trust privacy-preserving authentication model. Blockchain ensures that every user's identity and actions are securely validated before any model training begins, preventing malicious actors from accessing the system. Once the authentication is completed, the FL server assigns model training tasks to the computing nodes, which use blockchain to log the task assignments and track the training progress, ensuring transparency and trust in the training process.

Upon completion of model training, the nodes aggregate the trained models and send the fused results back to the users through a privacy and integrity-aware consensus-based fusion approach. The fusion process is secured by blockchain's immutable ledger, ensuring that the aggregation is transparent and that no unauthorized modifications can occur during the fusion process. Blockchain technology also guarantees that the model updates and aggregated results are auditable, further enhancing trust in the system.

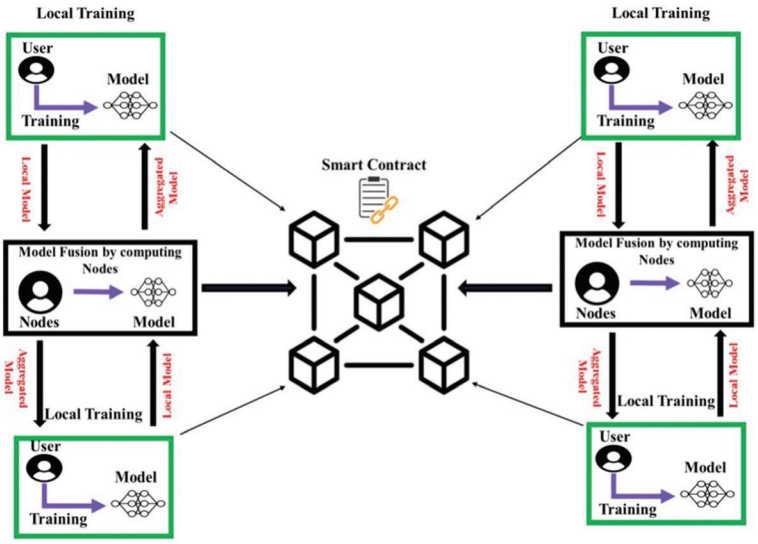


Fig. 1. Architecture of proposed Privacy-Integrity-Aware Blockchain Communication (PI- ABC) in federated edge-server learning platform

The PIABC approach, supported by blockchain's decentralized and secure infrastructure, ensures privacy and data integrity for each user without requiring any data sharing between users. To prevent potential security threats, such as inference or pollution attacks, a firewall is established between users. This firewall enforces a strict no-data-sharing policy while still allowing nodes to act as intermediaries for any necessary data transmission between users, thus ensuring enhanced security. In this setup, the nodes are responsible for executing model training and aggregation, while the FL server assigns the model training tasks. Blockchain provides a transparent and immutable record of task assignments, node activities, and model updates, ensuring the security and accountability of the entire system. The pseudocode of the same is presented in Algorithm 1.

3.2. BLOCKCHAIN-BASED TRUST PRIVACY PRESERVING AUTHENTICATION MODEL

The blockchain-based trust privacy-preserving authentication model ensures secure and verifiable interactions between users and computing nodes by utilizing dynamic trust evaluations combined with blockchain's immutable features. This model is designed to continuously assess and adjust trust levels between users and computing nodes based on a variety of metrics. These metrics include recent and past interactions, connection failures, and indirect trust derived from other nodes within the network, which are explained in further detail in the following sections. To achieve dynamic trust allocation, the model assigns trust weights to nodes based on their performance and reliability, allowing the system to identify and isolate malicious nodes effectively.

This approach guarantees that only trustworthy nodes are engaged in the federated learning process, maintaining the integrity of the system. Malicious or unreliable nodes are penalized, while trustworthy nodes are rewarded, incentivizing positive behavior and ensuring a secure and privacy-preserving environment for all users.

The trust levels are continually updated and validated through an exponential-average updating process, which ensures that the system remains resilient to manipulation and changes in node behavior. This ongoing validation of trust protects the system from fraudulent activities while fostering secure and reliable connections between users and nodes.

3.2.1. Direct and Indirect Trust Establishment

This work presents a validation and trustworthiness approach that authenticates the user and establishes secure interactions with computing nodes (workers). Initially, in this model, a trust level is calculated, which involves evaluating the trust a user has in a computing node (worker). It is important to note that each computing node stores relevant data, including user ID, time, data type, data size, and trust level, for the entire established connection between the user and the computing node.

To optimize storage overhead and efficiently manage the data collected by the computing node, this work allocates specific weights for storing the data. An exponential-average updating process is used to directly store this data into the Interplanetary File System (IPFS), ensuring scalability and decentralization.

Let $Sec_o^u(x, y)$,where x denotes user, y denotes computing node, o denotes total interaction time for given data-type, and u denotes time-period. The parameter Sec_o^u is used for evaluating trust-level between user and computing node. In this work, the direct trust is established between computing node y and user x and also between computing node y to computing node p and computing node p to user x.

The trust calculation process starts with an initial security metric. This initial value is typically set to a baseline value of $L_o^u = 0.5$ for any new interaction. This baseline represents neutral trust, implying that there is no prior information or bias about the trustworthiness of the computing node or user. Essentially, it establishes a starting point where neither trust nor distrust is assumed. The value of L_o^u is then updated dynamically based on the ongoing interactions between the user and the computing node, with more recent interactions having a greater influence on the trust level.

Direct trust is established in this work in three phases: between the user x and computing node y, between computing node y and computing node p, and between computing node p and user x. This connection is represented by the direct trust metric $L_o^u(x, y)$, which is mathematically expressed in Eq. (1).

$$\mathbb{L}_o^u(x, y) = Sec_o^u(x, y) \quad (1)$$

Algorithm 1 PIABC - Privacy-Integrity-Aware Blockchain Communication Framework

Require:

- User nodes $X = \{x_1, x_2, ..., x_n\}$ (perform local model training)
- Computing nodes $Y = \{y_1, y_2, ..., y_m\}$ (perform model aggregation)
- Federated Learning server FL
- Initial model M
- Blockchain ledger B
- Time period u, interaction duration o

Ensure: Aggregated global model M_{agg} delivered to authenticated users

Step 1: User Authentication
for each user $x_i \in X$ **do**
  Verify user x_i via blockchain-based authentication
  Log authentication event in blockchain ledger B
end for
Step 2: Model Assignment and Local Training
for each authenticated user x_i **do**
  Distribute initial model M to user x_i
  User x_i performs local training to generate model M_i
  Record training status and progress in blockchain B
end for
Step 3: Computing Node Selection
for each computing node $y_j \in C$ **do**
  Evaluate node y_j for aggregation eligibility
  if node y_j meets selection criteria **then**
    Assign aggregation task to node y_j
    Log task assignment in blockchain B
  else
    Mark node as untrustcd and log rejection in B
  end if
end for
Step 4: Model Aggregation
for each selected computing node y_j **do**
  Collect local models $\{M_1, M_2, M_n\}$ from user nodes
  Perform privacy-aware consensus-based fusion to generate M_{agg}
  Record aggregation process and final model hash in blockchain B
end for
Step 5: Secure Model Distribution
for each authenticated user x_i **do**
  Deliver final aggregated model M_{agg} to user x_i
  Log delivery event in blockchain ledger B
end for
  return M_{agg}

From Eq. (1), if computing node y provides better execution, then user x establishes connection having best trust-level. Similar happens with the computing node y if it gives better execution, then user x establishes connection with best trust-level. This helps user x to achieve direct trust establishment.

In this work, the indirect trust is established between the computing node y and user x and also between computing node y to computing node p and computing node p to user x by considering past established connection. To gain knowledge about past established connections, the computing node y connects with computing nodes p to collect the trust-level previously established by user x. Finally, computing node y fuses (aggregates) trust-level established from computing nodes p using Eq. (2).

$$\mathbb{G}_o^u(x,y) = \begin{cases} \frac{\sum_{p \in Z-\{x\}} \mathbb{F}_o^u(x,p) * \mathbb{L}_o^u(x,y)}{\sum_{p \in Z-\{x\}} \mathbb{F}_o^u(x,p)}, if\ |Z-\{x\}| > 0 \\ 0, if\ |Z-\{x\}| = 0 \end{cases} \quad (2)$$

In Eq. (2), $G_o^u(x, y)$ denotes fused trust-levels collected from computing nodes p and $Z=S(y)$ denotes computing node p which had established connection with computing node y.

This work utilizes weight-based approach, where weights are allocated dynamically for computing nodes. For allocation, higher weights are allocated for highly-trusted computing nodes, whereas lower weights are allocated for less-trusted computing nodes. Consider $F_o^u(x, y)$ which is used to denote the evaluation of validation-based security trustworthiness for a computing node y. The F_o^u (x,y) is mathematically evaluated using Eq. (3).

$$\mathbb{F}_o^u(x,y) = \begin{cases} 1 - \frac{\log\left(Sec_o^u(x,y)\right)}{\log\theta}, & if\ \mathbb{R}_o^u(x,y) > \theta, \\ 0, & else \end{cases} \quad (3)$$

In Eq. (3), logθ denotes similar least-tolerable variable and $R_o^u(x, y)$ denotes relationship I between user x and other computing nodes y (where $p=y$ for one established connection between worker and computing node).

3.2.2. Evaluation of Latest and Past Established Trust

In this work, the latest trust is established between computing node y and user x and also between computing node y to computing node p and computing node p to user x is evaluated by considering both direct and indirect-trust. During evaluation of latest established trust, direct established trust is given higher trust-level, because the computing node y or computing node p interacts more with the user x. Hence, the latest established trust is mathematically evaluated using Eq. (4).

$$\mathbb{C}_o^u(x,y) = \delta * \mathbb{L}_o^u(x,y) + (1-\delta) * \mathbb{G}_o^u(x,y) \quad (4)$$

In Eq. (4), $C_o^u(x, y)$ denotes latest established trust metric and δ denotes trust-level weight assigned to direct established trust. δ is weighted function that can be optimized dynamically according to users x interaction $T^u(x, y)$ on respective worker nodes y considering time u; however, in this work average interaction $\vec{T}^u(x, y)$ time is considered to dynamically optimize the δ weighted value.

In Eq. (4), as u increases, the latest establish trust becomes old, which can be termed as past established trust and is denoted as $Q_o^u(x, y)$. The evaluation of $Q_o^u(x, y)$ is done similar to trust-level evaluation, i.e., using EAUP. Hence $Q_o^u(x, y)$ can be mathematically evaluated using Eq. (5).

$$\mathbb{Q}_o^u(x,y) = \frac{\varphi * \mathbb{L}_{o-1}^u(x,y) + \mathbb{C}_{o-1}^u(x,y)}{2} \quad (5)$$

In Eq. (5), φ(0≤φ≤1) is the incentive parameter and whenever L_0^0 (x,y), the whole evaluation changes to 0. By utilizing past established trust, the malicious computing nodes y connecting with computing nodes p or user x cannot change their process, i.e., computing node p cannot connect with y or x when a y has already established a connection with x. Hence, this metric allows a computing node y or computing node p to establish a connection with user x in a cooperative way, thereby reducing attacks. Also, the trust-level for the latest established trust changes to past established trust only when a computing node y or computing node p has made more connections with user x, hence increasing privacy.

3.2.3. Evaluation of Upcoming Trust Establishment

In this work, the upcoming trust is established between computing node y and user x and also between computing node y to computing node p and computing node p to user x is evaluated by considering both latest and past established trust. Consider $Future_o^u(x, y)$ as upcoming trust which will be established from computing node y to user x can be mathematically represented using Eq. (6).

$$Future_o^u(x,y) = \begin{cases} 0, if\ neither\ \mathbb{Q}\ or\ \mathbb{C}\ is\ available \\ \alpha\mathbb{C}_o^u(x,y) + (1-\alpha)\mathbb{L}_o^u(x,y)\ if\ either\ \mathbb{Q}\ or\ \mathbb{C}\ is\ available \end{cases} \quad (6)$$

In Eq. (6), α is a dynamic variable and in this work α=0. The α can be dynamically changed using a deviating variable ω according to application/task requirement. Also, by making ω dynamic, a computing node can change its past established trust to latest established trust. Moreover, it is important that ω should not be set very less as malicious computing nodes can use this parameter for changing their behavior, i.e., they may change from malicious to non-malicious computing node, hence leading to attack to user x.

3.2.4. Security Metric for classification of Malicious Computing Node

For identifying and classifying malicious computing node, this work considers a security metric denoted as $F_o^u(x, y)$, which is evaluated by considering upcoming trust establishment and unfair and changing trust metric, which is mathematically represented using Eq. (7).

$$\mathcal{F}_o^u(x,y) = \mathbb{Q}_o^u(x,y) * Future_o^u(x,y) \quad (7)$$

Using Eq. (7), the computing nodes having higher upcoming trust-level, will result in less unfair and changing trust-level. Hence, the malicious computing nodes will have lesser trust-level, thereby reducing the attack on user x. Also, to having higher trust-levels during upcoming trust establishment, it is necessary that it should not change its process. Hence, using Eq. (7), user x can select the computing node y having higher trust-level, thereby reducing attack and increasing security.

3.3. PRIVACY AND INTEGRITY-AWARE CONSENSUS-BASED FUSION APPROACH

This section provides an efficient approach for fusing (aggregating) data to provide security, privacy, confidentiality and integrity. The aggregation process takes place after ensuring trust-level security and authenticating data privacy. Moreover, for providing integrity for authenticated data, a consensus-based fusion approach is presented, where the data which is unsecured is discarded

3.3.1. Privacy-Integrity Consensus Approach

This work provides integrity by using a consensus-based fusion approach to prevent diverse attack which includes pollution and inference attack for federated-learning environment. Consider a blockchain-based federated-learning environment, where there exists x users. By using graph-theory, this work considers users x as a graph H={E, V}, where H is a graph consisting of edges E (set of connections or interactions between users) and vertices V (set of users). Further, consider (j, k)∈E (which implies a directed interaction from vertex j to vertex k, meaning that user j sends data or model updates to user k), only if users are interconnected with each other. Consider the starting state of users as y_j (0), having different time-session l. As the user j will have contact or might establish a connection with other users.

The drawback of current consensus-based security approaches is that users in federated-learning environment try to get knowledge of other users starting states, i.e., y(0), hence impacting other user's privacy. The proposed consensus model during integrity assurance makes sure it preserves privacy requirements. Thus, the proposed consensus-based fusion approach utilizes $y(l)$ to converge to $\bar{y}$, for preserving user's privacy. The process of convergence is done in repetitive four steps so that it converges to $\bar{y}$. The steps are given below:

1. Assumption of Random Data: Consider that each user communicates random data w_j (l) for time-session l, where variance=1 and mean=0. Also consider that w_j (l) for various users is represented as $\{w_j(l)\}_{j=1,\ldots,\ o,\ l=0,\ 1,\ldots}$ which is uniformly distributed. For each user generating data and communication, a noise can be induced for every y_j (l) which can be represented using x_j (l). The noise x_j (l) can be denoted using Eq. (8).

$$x_j(l) = w_j(o) \tag{8}$$

2. Evaluation of Noise: The Eq. (8) is only correct whenever l=0, else x_j (l) is evaluated using Eq. (9).

$$x_j(l) = \beta^l w_j(l) - \beta^{l-1} w_j(l-1) \tag{9}$$

In Eq. (9), β is constant variable for every user and changes from 0 to 1. The novel state for new user can be obtained using Eq. (10).

$$y_j''(l+1) = b_{jj} y_j''(l) + x_j(l) \tag{10}$$

3. Interaction with Adjacent Users: Further, the users interacting with nearby (adjacent) users and their state mean is evaluated using Eq. (11).

$$y_j(l+1) = b_{jj} y_j''(l) + \sum_{k \in \mathcal{O}(j)} b_{jk} y_j''(l) \tag{11}$$

4. Updating States: Revise l+1 states and return back to Step 1. This process is repeated until $y(l)$ converges to $\bar{y}$. Further, Eq. (10) and Eq. (11) are converted to matrix format as presented in Eq. (12).

$$y(l+1) = By''(l) = B\big(y(l) + x(l)\big) \tag{12}$$

By utilizing the 4 Steps, the $y(l)$ converges to $\bar{y}$ having accurate average state values. Also, during convergence it is important to assure that noise too reduces. Moreover, to achieve best convergence result, the asymptotic-sum has to be 0. The presented consensus-based security approach provides better outcome for both general and gaussian noise by utilizing highest-probability evaluation method. Also, the consensus-based security approach does not require any communication-channel, hence, utilizes less energy and resources and in federated-learning environment. The presented consensus-based fusion approach also provides privacy and integrity as it need less values for constructing consensus as users can optimize noise $x(l)$ independently, rather than depending on any variable. Moreover, it is important to know that each user or multiple users can choose $x(l)$ independently with considering the exponential-decaying co-variance matrix, but has to ensure that $x(l)$ does not affect consensus variables to achieve correct average consensus.

The data fused using Eq. (12) has better trust-level security privacy and integrity, but verifying data is important, hence, it is important to identify if there is any malicious packet present in network or not. For this, consider J_0 where data is fused for achieving better security to prevent attacks and J_1 denotes non-efficient data where attacks could happen. Hence from this, the attack can be identified by considering non-malicious packets interaction. A non-malicious packet will show no changes in state, whereas malicious packet will show a change. The state of normal packet can be denoted as R_h=R ($J_1 \mid J_0$) and malicious packet can be represented as R_m=R ($J_0 \mid J_1$). Hence a static-test can be developed for this evaluation, where initially the data variance is evaluated using Eq. (13).

$$N = \| \, \mathrm{y}_j(l) - \hat{y}_j(l) \, \|^2 \tag{13}$$

Eq. (13), provides difference among original data and original + noise data. For identifying malicious packet and non-malicious packets, the variable $N \lessgtr_{J_1}^{J_0} (\vartheta)$ is used. If there exists malicious packet, then the users can be changed to malicious, else they remain normal. Also, it has to be noted that the malicious packets consume more resources as they try to attack other nodes/packets; this helps in identifying the probability of attack within the federated learning environment. The proposed use of both trust-based authentication combined with a consensus-based privacy-integrity assurance model is effective in authenticating the user and eliminating false packets within the federated edge-server learning environment; the process improves overall throughput with higher detection accuracy and less misclassification, as proved in the following section.

3.3.2. Theoretical Analysis and Security Guarantees

The previous section outlined the equations for trust computation and consensus-based privacy–integrity fusion. In this section, we present theoretical analyses and bounds demonstrating the model's behavior across different scenarios.

Convergence Analysis of Trust Evaluation: We provide a mathematical discussion showing that the exponential-average updating process used for both trust estimation and consensus fusion converges to a stable value over time, assuming bounded variance in interaction behavior. A formal convergence guarantee has been added by analyzing the recursive structure of Equations (4), (5), and (6) using principles from Markov decision processes and stochastic averaging.

Security Proof of Malicious Node Isolation: We analytically derive that malicious nodes characterized by fluctuating or degrading trust values are penalized in successive rounds due to diminishing trust weights (as shown in Equation (7)). This ensures that their influence on the federated learning process is minimized, and they are progressively isolated.

Consensus Robustness under Noise and Attack: We added a theoretical explanation of how the consensus protocol resists data pollution and inference attacks. By modeling user interaction as a graph and the injected noise as a bounded Gaussian variable, we show that the mean state converges with high probability to the correct average (Equation 12), supported by an asymptotic variance reduction analysis.

Formal Attack Detection Bound: For the malicious data detection metric (Equation 13), we now provide a statistical hypothesis testing model based on the Neyman-Pearson Lemma to distinguish between hypotheses J_0 (benign) and J_1 (malicious). This includes specifying decision thresholds ϑ and false positive/negative rates.

4. RESULTS AND DISCUSSION

To validate the practical applicability of the proposed PIABC security model, we simulate its deployment in a federated IoT environment under varying adversarial conditions. The objective is to assess the model's ability to detect and isolate malicious participants effectively while maintaining low misclassification rates and high throughput. PIABC's performance is benchmarked against an existing blockchain-aided privacy-preserving framework (BPPF) [18], using both synthetic and real-world datasets. Simulation scenarios are designed to reflect dynamic attack intensities, diverse threat types, and realistic network conditions, as detailed in the following section.

4.1. SIMULATION SCENARIO FOR REALISTIC DEPLOYMENT

Experiments were conducted to evaluate the PIABC model against the blockchain-aided privacy-preserving framework (BPPF) [18] using the CIC Federated Learning Dataset [19]. Performance was measured in terms of detection rate, misclassification rate, and throughput under attack intensities ranging from 10% to 40%. Both models were implemented in the C#-based SENSORIA simulator [20] with blockchain support via IoTSim-Osmosis [21].

To further evaluate the robustness of the PIABC model, simulations were performed on the UNSW-NB15 and CIC-IoT2023 datasets [20], which contain diverse multi-stage cyberattacks. Sybil and collusion attacks are mitigated through cross-node trust validation and anomaly detection, while backdoor threats are countered by tracking temporal trust variations and applying consensus-based integrity verification.

Before training, both datasets were preprocessed using a standard pipeline: removal of duplicate and incomplete records, one-hot encoding for categorical variables, and Min-Max normalization for numerical attributes. Highly correlated and low-variance features were filtered out, and final feature subsets were chosen based on statistical relevance and domain knowledge—25 features from UNSW-NB15 and 28 from CIC-IoT2023. The processed data was then split into training (70%), validation (15%), and testing (15%) sets using stratified sampling to maintain class balance. A fixed random seed ensured reproducibility, and 5-fold cross-validation was employed for hyperparameter optimization and anomaly threshold selection.

The simulation results demonstrate that PIABC consistently achieves higher detection accuracy and lower misclassification rates across varied attack scenarios, confirming its effectiveness and scalability in real-world federated IoT deployments.

4.2. DETECTION RATE

This section studies the detection rate performance of identifying the attack using both PIABC and BPPF under

the same simulation configuration. A higher value of detection rate indicates superior performance. The detection rate performance of both models is graphically shown in Figure 2. The result shows the proposed PIABC model has a higher detection rate in identifying the attack in comparison with BPPF, considering varied attack percentages. The enhancement achieved is due to the adoption of an effective trust model implemented in Eq. (7).

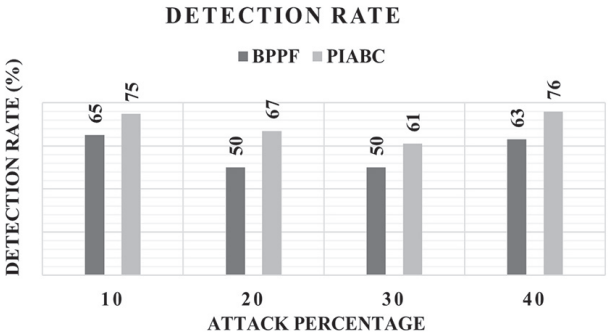


Fig. 2. Detection rate vs varied attack percentage

4.3. MISCLASSIFICATION RATE

This section studies the misclassification rate performance of wrongly identifying the attack using both PIABC and BPPF under the same simulation configuration. A lower value of the misclassification rate indicates superior performance. The misclassification rate performance of both models is graphically shown in Figure 3. The result shows the proposed PIABC model has a higher misclassification rate in wrongly identifying the attack in comparison with BPPF, considering varied attack percentages. The enhancement achieved is due to the adoption of an effective consensus model designed in Eq. (12).

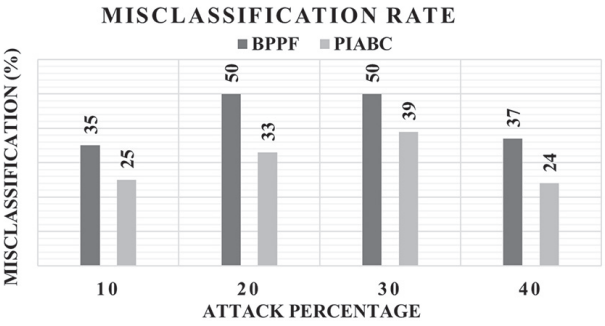


Fig. 3. Misclassification rate vs varied attack percentage

4.4. NORMALIZED THROUGHPUT

This section studies the normalized throughput performance by varying attack rate using both PIABC and BPPF under the same simulation configuration. The throughput is measured in terms of bits transmitted per second; however, in this research article normalized throughput is considered for validation.

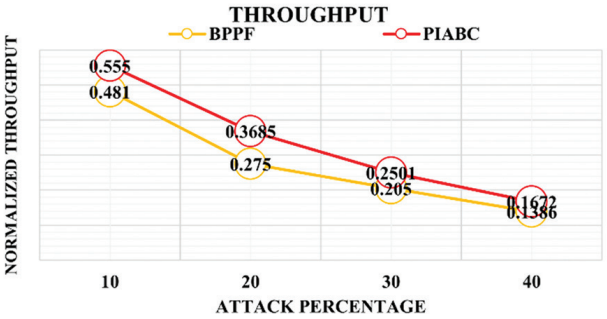


Fig. 4. Normalized throughput vs varied attack percentage

A higher value of normalized throughput indicates superior performance. The detection rate performance of both the models is graphically shown in Fig. 4. The result shows the proposed PIABC model has higher normalized throughput in comparison with BPPF considering varied attack percentage. The normalized throughput enhancement achieved is due to adoption of effective trust model implemented in Eq. (7) and consensus model designed in Eq. (12).

4.5. DISCUSSION

The PIABC attains convergence, by providing a formal proof showing that the exponential-average updating process converges under bounded interaction variance. Regarding security, we use statistical hypothesis testing to demonstrate resilience against Sybil and backdoor attacks based on trust deviation detection. For computational complexity, we analytically derive that the trust evaluation and consensus fusion algorithms operate with polynomial time complexity $O(n \cdot t)$, where n is the number of nodes and t is interaction time, ensuring scalability. The PIABC system employs a Proof-of-Authority (PoA) consensus protocol, integrated via the IoTSim-Osmosis framework, due to its lightweight nature and suitability for resource-constrained IoT environments. PoA allows for faster block confirmations and lower energy consumption compared to Proof-of-Work, making it ideal for real-time intrusion detection in federated learning-based systems. Blockchain-induced delays, including block generation and smart contract execution times, are simulated using event-driven modeling within SENSORIA, incorporating realistic network latency based on exponential distribution patterns. The storage model is designed for efficiency, with only hash-verified metadata, such as access logs and model update references, stored on-chain, while bulk data remains securely stored off-chain in cloud repositories. This hybrid approach ensures transparency, data integrity, and scalability. The measured computational overhead of the blockchain integration remains under 5%, while communication overhead is limited to 7–8%, primarily due to compact transaction sizes (~256 bytes). These overheads are directly correlated with key performance metrics. Despite the additional cost and processing load, the proposed PIABC model achieves

improved throughput, higher detection rates, and reduced misclassification rates when compared with the baseline BPPF framework. The reduction in false positives and increased detection accuracy justifies the minimal added overhead, while delay remains within acceptable thresholds.

5. CONCLUSION

This work shows that in federated learning, the data model is executed in multiple edge-server in a collaborative manner; as a result, it increases users' privacy and data breach because of a single point failure attack in the main computational server. Blockchain employing rewarding mechanism in a federated edge-learning platform aids the model to overcome single-point aggregation failure. However, the current method failed to identify selfish and biased workers; further, reaching global consensus model to assure privacy-integrity in blockchain-enabled federated edge-server is difficult. This work introduced privacy-integrity-aware blockchain communication (PIABC) in federated edge-server learning platform. An experiment is conducted to study the performance considering a varied attack size. The results show the proposed model can resist different attacks using the CIC-IOT federated edge attack dataset. The average percentage reduction in detection rate by PIABC over BPPF is approximately 18.46%. The average percentage improvement in misclassification rate by PIABC over BPPF is approximately 26.13%. The average percentage improvement in throughput by PIABC over BPPF is approximately 44.53%.

The PIABC model is very effective in comparison with the existing blockchain-privacy preserving scheme for identifying the correctly aggregated packets and eliminating malicious packets within the federated edge-server learning platform. Future work would consider developing more effective detection strategies to detect more complex attacks employing different benchmarks and also further optimizing the model. In the current work, we conducted detailed simulations to evaluate throughput, detection rate, misclassification rate, and computational complexity to measure the overhead of the proposed model. These evaluations demonstrate the model's effectiveness and low processing cost in simulated environments. However, aspects such as energy consumption, real-time latency, and deployment on actual edge hardware have not been covered in this study. Therefore, the future research will aim to validate the model's effectiveness and resource efficiency through deployment on actual edge computing platforms.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support provided by the Gujarat Council on Science and Technology (GUJCOST), India, for the research project, under project number GUJCOST/STI/2021-22/3867. This support has been instrumental in the successful completion of this research work.

6. REFERENCES:

[1] H. Li, L. Ge, L. Tian, "Survey: federated learning data security and privacy-preserving in edge-Internet of Things", Artificial Intelligence Review, Vol. 57, No. 5, 2024, p. 130.

[2] T. Alam, R. Gupta, A Ullah, S. Qamar, "Blockchain-Enabled Federated Reinforcement Learning (B-FRL) model for privacy preservation service in IoT systems", Wireless Personal Communications, Vol. 136, No. 4, 2024, pp. 2545-2571.

[3] Y. Jia, L. Xiong, Y. Fan, W. Liang, N. Xiong, F. Xiao, "Blockchain-based privacy- preserving multi-tasks federated learning framework", Connection Science, Vol. 36, No. 1, 2024, p. 2299103.

[4] J. Shen, S. Zhou, F. Xiao, "Research on Data Quality Governance for Federated Cooperation Scenarios", Electronics, Vol. 13, No. 18, 2024, p. 3606.

[5] K. M. Sameera, S. Nicolazzo, M. Arazzi, A. Nocera, R. R. KA, P. Vinod, M. Conti," Privacy-preserving in Blockchain-based Federated Learning systems", Computer Communications, Vol. 222, 2024, pp. 38-67.

[6] C. Dhasaratha, M. K. Hasan, S. Islam, S. Khapre, S. Abdullah, T. M. Ghazal, A. I. Alzahrani, N. Alalwan, N. Vo, M. Akhtaruzzaman, "Data privacy model using blockchain reinforcement federated learning approach for scalable internet of medical things", CAAI Transactions on Intelligence Technology, 2024. (in press)

[7] Z. Jovanovic, Z. Hou, K. Biswas, V. Muthukkumarasamy, "Robust integration of blockchain and explainable federated learning for automated credit scoring", Computer Networks, Vol. 243, 2024, p. 110303.

[8] S. T. Ahmed, T. R. Mahesh, E. Srividhya, V. Vinoth Kumar, S. B. Khan, A. Albuali, A. Almusharraf, "Towards blockchain based federated learning in categorizing healthcare monitoring devices on artificial intelligence of medical things investigative framework", BMC Medical Imaging Vol. 24, No. 1, 2024, p. 105.

[9] W. Moulahi, I. Jdey, T. Moulahi, M. Alawida, A. Alabdulatif," A blockchain-based federated learning mechanism for privacy preservation of healthcare

IoT data", Computers in Biology and Medicine, Vol. 167, 2023, p. 107630.

[10] L. Wang, C. Guan, "Improving security in the internet of vehicles: A blockchain-based data sharing scheme", Electronics, Vol. 13, No. 4, 2024, p. 714.

[11] H. Liu, H. Zhou, H. Chen, Y. Yan, J. Huang, A. Xiong, S. Yang, J. Chen, S. Guo, "A federated learning multi-task scheduling mechanism based on trusted computing sandbox", Sensors, Vol. 23, No. 4, 2023, p. 2093.

[12] W. E. Mbonu, C. Maple, G. Epiphaniou, "An end-process blockchain-based secure aggregation mechanism using Federated Machine Learning", Electronics, Vol. 12, No. 21, 2023, p. 4543.

[13] M. A. Mohammed, A. Lakhan, K. H. Abdulkareem, D. A. Zebari, J. Nedoma, R. Martinek, S. Kadry, B. Garcia-Zapirain, "Energy-efficient distributed federated learning offloading and scheduling healthcare system in blockchain based networks", Internet of Things, Vol. 22, 2023, p. 100815.

[14] M. Zirui, G. Bin, "A Privacy-Preserved and User Self-Governance Blockchain-Based Framework to Combat COVID-19 Depression in social media", IEEE Access, Vol. 11, 2023, pp. 35255-35280.

[15] H. Javed, Z. Abaid, S. Akbar, K. Ullah, A. Ahmad, A. Saeed, H. Ali, Y. Y. Ghadi, T. J. Alahmadi, H. K. Alkahtani, A. Raza, "Blockchain-based logging to defeat malicious insiders: The case of remote health monitoring systems", IEEE Access, Vol. 12, 2023, pp. 12062-12079.

[16] I. U. Din, K. A. Awan, A. Almogren, "Secure and privacy-preserving trust management system for trustworthy communications in intelligent transportation systems", IEEE Access, Vol. 11, pp. 65407-65417.

[17] Q. Xie, Z. Sun, Q. Xie, Z. Ding," A cross-trusted authority authentication protocol for Internet of Vehicles based on blockchain", IEEE Access, Vol. 11, 2023, pp. 97840-97851.

[18] Z. Ma, J. Jiang, H. Wei, B. Wang, W. Luo, H. Luo, D. Liu, "A blockchain-based secure distributed authentication scheme for internet of vehicles", IEEE Access, Vol. 12, 2024, p. 81471-81482.

[19] Datasets — Research — Canadian Institute for Cybersecurity — UNB, https://www.unb.ca/cic/datasets/index.html (accessed: 2024)

[20] N. Ababneh, J. N. Al-Karaki, "On the lifetime analytics of IoT network", Proceedings of the International Conference on Communication and Signal Processing, Chennai, India, 28-30 July 2020, pp. 1086-1090.

[21] A. Albshri, A. Alzubaidi, M. Alharby, B. Awaji, K. Mitra, E. Solaiman, "A conceptual architecture for simulating blockchain-based IoT ecosystems", Journal of Cloud Computing, Vol. 12, No. 1, 2023, p. 103.

Parallel and Distributed Multi-level Entropy-Based Approach for Adaptive Global Frequent Pattern Mining in Large Datasets

Original Scientific Paper

Houda Essalmi*
Laboratory of Engineering Sciences, Polydisciplinary Faculty of Taza,
University of Sidi Mohamed Ben Abdellah
Fez, Morocco
houda.essalmi@usmba.ac.ma

Anass El Affar
Laboratory of Engineering Sciences, Polydisciplinary Faculty of Taza,
University of Sidi Mohamed Ben Abdellah
Fez, Morocco
anass.elaffar@usmba.ac.ma

*Corresponding author

Abstract *– Frequent pattern mining in distributed settings remains a significant challenge due to predominantly high computational expenses and high communication overhead. This paper presents AGFPM (Adaptive Global Frequent Pattern Mining), a novel solution that integrates an extensible Master-Slave architecture with an advanced pruning technique that relies on binary entropy and statistical quartiles. AGFPM proposes two primary data structures: the LP-Tree (Local Prefix Tree) and the GP-Tree (Global Prefix Tree). A single pass of each local Slave site is used to build one LP-Tree, and low information value branches are pruned early on by entropy and quartile thresholds. Rather than transferring complete trees, only succinct metadata is sent to the Master site, where the GP-Tree is built from globally sorted items in order of their entropy rankings. A significant aspect of AGFPM is the flexible pruning approach: either the GP-Tree is pruned or not pruned, based on user criteria. This provides a dynamic adjustment between the performance and generality of results, thereby allowing control over the level of compression applied when generating global patterns. Global frequent patterns are then recursively mined from the GP-Tree based on conditional sub-GP-Trees. Frequent patterns are extended at each level of the hierarchy by intersecting the common prefix paths, guided by a Global Header Table. AGFM demonstrates improved performance in execution time, scalability, and robustness against low support thresholds relative to existing methods.*

Keywords: *Data mining, Distributed Datasets, FP-tree; Communication Overhead, Frequent patterns mining, Binary Entropy, Quartile-based Pruning*

Received: July 2, 2025; Received in revised form: August 30, 2025; Accepted: September 1, 2025

1. INTRODUCTION

1.1. BACKGROUND AND CHALLENGES

Frequent pattern mining over distributed, horizontally partitioned data aims to uncover meaningful co-occurrences while keeping computation and communication overheads manageable [1-3]. Candidate-generation approaches rooted in the Apriori principle typically face a rapid growth in candidates and require multiple scans of the data. Pattern-growth methods mitigate this by compressing transactions into compact prefix structures, yet they can still encounter memory pressure and expensive cross-site consolidation as datasets become large or dense [4-7]. In practice, an effective solution should combine compact and merge-friendly structures, a coherent global ordering that simplifies fusion, limited network traffic, and principled filtering to constrain structural growth without discarding informative patterns.

1.2. LIMITS OF PRIOR APPROACHES

Algorithms in the Apriori family rely on an iterative process of candidate generation and support validation, guided by the downward closure property [8]. While effective, this approach necessitates multiple database

scans and extensive synchronization, leading to scalability challenges in distributed environments [8-12]. Several distributed variants, such as Count Distribution (CD) [9], Distributed Mining Algorithm (DMA) [10], and Fast Distributed Mining (FDM) [11], operate by aggregating local counts across nodes, but often suffer from candidate set explosion and high communication overhead, especially at low minimum support thresholds. Enhanced methods like Optimized Distributed Association Mining (ODAM) [12] and Distributed Decision Miner (DDM) [13] aim to mitigate these issues by reducing inter-site message exchange and memory consumption. To improve efficiency, parallelization strategies leverage data or task decomposition, including Data Distribution (DD) [9], Intelligent Data Distribution (IDD) [14], and Hash-based Parallel Association (HPA) [15], which distribute workloads to increase throughput. Techniques such as Scalable Hybrid (SH) [16] address load imbalance by adapting to data skew. Hybrid approaches like Candidate Distribution (CaD) [9] and Hybrid Distribution (HD) [14] further combine partitioning strategies using prefix-based assignment or optimized memory-aware distribution to minimize communication and enhance locality. Despite these optimizations, the fundamental limitations of repeated data scans and coordination between nodes continue to hinder performance in large-scale or dense datasets.

FP-growth avoids explicit candidate generation by organizing transactions into a compact prefix tree and mining conditional pattern bases. Parallel variants such as PFP-tree and Multiple Local Parallel Trees (MLPT) construct local trees across sites and merge them subsequently, alleviating candidate-generation costs but introducing heavy cross-site transfers and structural heterogeneity that complicate global consolidation [17, 18]. Single-pass and compact prefix structures reduce scans and memory requirements; however, harmonizing many sizable local trees under constrained memory and bandwidth remains challenging [4-6].

In dense settings and at low support, both candidate-generation and pattern-growth families face similar stressors: intermediate structures can grow quickly (large candidate sets or deep/wide trees), local structures diverge across sites and become difficult to align, and communication scales with the number of sites, especially during counting and merging [4–16]. Because control is predominantly frequency-based, branches with limited informational value often persist, inflating consolidation costs and post-processing effort. These observations point to the need for mechanisms that are explicitly information-aware and distribution-adaptive, so that uninformative regions are pruned early, local structures are better aligned before fusion, and global aggregation is streamlined.

1.3. MOTIVATION: BINARY ENTROPY WITH QUARTILE-BASED PRUNING

Fixed frequency thresholds (e.g., MinSupp) do not always provide sufficient control over structural growth and redundancy in distributed trees. Binary Shannon entropy offers an information-theoretic signal of presence/absence uncertainty for items and paths, distinguishing informative from low-value branches [19, 20]. Using the empirical distribution of entropies, adaptive quartile thresholds (Q1, Q2, Q3) enable progressive pruning: rapid removal of minimally informative branches (below Q1), complementary reduction around the median (Q2), and preservation of the most information-rich parts (above Q3) [21, 22]. This data-adaptive strategy avoids arbitrary hyperparameters and can be applied locally to shrink LP-Trees prior to fusion and, optionally, globally to regulate GP-Tree complexity.

1.4. CONTRIBUTIONS OF AGFPM

This work presents AGFPM (Adaptive Global Frequent Pattern Mining), a distributed method built around four complementary pillars. First, it adopts a Master–Slave architecture with lightweight communication: each site constructs a (Local Prefix-Tree) LP-Tree and shares only compact metadata with the Master, thereby avoiding full-tree transfers and reducing network overhead. Second, it imposes a global, information-driven ordering: items are ranked by global entropy, and sites rebuild their LP-Trees accordingly, yielding structurally aligned trees that facilitate accurate and efficient fusion. Third, it implements multi-level adaptive pruning: branches are filtered using binary entropy and quartile thresholds (Q1, Q2, Q3), applied locally to the (Global Prefix-Tree) GP-Tree, in order to contain structural growth, reduce redundancy, and preserve the most informative paths. Fourth, it enables efficient global mining: a GP-Tree, guided by a Global Header Table and the global entropy order, supports recursive, top-down extraction via conditional sub-GP-Trees. Collectively, these design choices shorten runtime and communication, improve robustness under low support on dense datasets, and produce more focused sets of global patterns without compromising relevance findings corroborated on standard synthetic and real benchmarks [23-25].

Positioned against prior work, AGFPM extends beyond Apriori-style distributed families [8–16] by dispensing with explicit candidate generation and limiting synchronization, and beyond parallel pattern-growth approaches [4-7, 17-18] by transmitting succinct metadata and enforcing a global entropy-based order that homogenizes LP-Trees before fusion. In addition, AGFPM introduces a multi-level, quartile-guided pruning strategy applied locally and, when desired, to the GP-Tree to bound structural complexity and curb redundancy. Relative to recent distributed big-data systems [26-31], the combination of information-driven ordering and adaptive pruning jointly lowers communication overhead and runtime while strengthening robustness at low support on dense data.

The remainder of this paper is structured as follows: Section 2 presents related work. Section 3 discusses the problem definition of mining frequent patterns in

parallel and distributed contexts and presents the concept of Entropy and Quartiles. The proposed algorithm is given in detail in Section 4. The results and discussion are presented in Section 5. Finally, Section 5 draws conclusions from this work.

2. RELATED WORK

Several studies have been suggested to improve the way pattern mining algorithms process large data. We discuss important contributions in distributed frequent pattern mining in this section.

Deng *et al*. [26] proposed an optimized and distributed version of the Apriori algorithm, titled STB_Apriori, designed to operate on the Apache Spark platform. This approach stands out due to the use of BitSet structures, which allow transactions to be represented in a compressed and efficient manner, significantly reducing the memory required for processing. By fully leveraging Spark's in-memory computing model, the algorithm manages to limit disk accesses while benefiting from distributed parallelism across data partitions. The authors demonstrated, through experiments on large datasets, that STB_Apriori outperforms the classic versions of Apriori, both in terms of execution time and scalability. These results confirm the value of combining memory optimization techniques with distributed frameworks to improve the extraction of frequent itemsets in large-scale environments.

Shaikh *et al*. [27] developed DIAFM (Distributed Incremental Approximation Frequent Itemset Mining), a distributed approach designed for the incremental extraction of frequent itemsets at a large scale. This algorithm relies on the MapReduce model and introduces a strategy of successive fragment processing, allowing for the efficient integration of new data without re-scanning the entire dataset. DIAFM aims to reduce the overall computational cost while maintaining high accuracy, thanks to a controlled error tolerance mechanism. Experimental evaluations show that this method significantly improves processing time while adapting to dynamic environments such as transactional streams or continuous monitoring systems. Its modular architecture and compatibility with increasing data volumes make it a promising solution for incremental exploration in distributed contexts.

Sun *et al*. [28] developed a distributed basket analysis system for large transaction volumes, leveraging the parallel processing capabilities of Apache Spark. This approach relies on a two-step execution. First, random portions of the data are analyzed locally using FP-Growth to identify partial frequent patterns. In a second step, these results are consolidated using an approximate method to construct a global set of representative itemsets. The system stands out for its flexibility, particularly due to the integration of Spark SQL, which facilitates the analytical querying of the extracted results. Experiments conducted on massive datasets (up to one billion transactions) demonstrate a significant reduction in execution time and excellent scalability, while maintaining accuracy close to that of exact methods.

In a recent study, Raj *et al*. [29] proposed CrossFIM, a hybrid algorithm for large-scale distributed extraction of frequent itemsets, specifically designed for the Apache Spark framework. This method dynamically alternates between the Apriori and Eclat strategies depending on the depth of the analysis, in order to combine the efficiency of candidate generation with the speed of tidset intersections. Particular attention was paid to optimizing communications between nodes through a clever partitioning mechanism, significantly reducing the volume of exchanged data. Experimental results indicate that CrossFIM offers excellent scalability and efficient memory resource management, making it particularly suitable for analyzing very large transactional datasets.

Rochd and Hafidi [30] introduce DSSS (Distributed Single Scan on Spark), an algorithm designed to optimize the extraction of frequent itemsets in a Big Data environment, fully leveraging the distributed capabilities of Apache Spark. Their method relies on a single scan strategy of the transactional database, which contrasts sharply with traditional algorithms requiring multiple successive passes. To achieve this, DSSS uses in-memory RDDs and a global hash table broadcast to all nodes. This design effectively reduces the latency caused by disk I/O and inter-node communications, which are often bottlenecks in distributed architectures. The algorithm also incorporates an optimization of the management of local support structures, dynamically filtering out irrelevant items before the aggregation phase. Experiments conducted on public datasets (notably Retail, T40I10D100K) show that DSSS outperforms several reference distributed algorithms, such as PFP-Tree, and D-Apriori, particularly on large and sparsely dense databases. In terms of scalability, the results indicate a significant reduction in execution time as the number of nodes increases, with a nearly linear behavior, demonstrating a good distribution of the processing load. Moreover, the algorithmic complexity remains manageable, as it primarily depends on the average number of items per transaction rather than the total size of the dataset.

Singla and Gandhi [31] propose an algorithm titled PDReLim ("Parallel and Distributed Recursive Elimination"). This work aims to optimize the mining of frequent itemsets in distributed environments, leveraging PySpark and Spark's RDD capabilities. PDReLim differs from traditional approaches through a recursive local pruning strategy: each node removes items deemed infrequent before generating new candidates, thereby significantly reducing the search space. This local optimization allows for limiting inter-node exchanges and controlling combinatorial complexity. The experiments, conducted on standard datasets (Chess, Mushroom, Connect) from the UCI repository, demonstrate that PDReLim significantly outperforms traditional iterative MapReduce methods such as PApriori, PFP-Growth, and PFP-Max, particularly at low support thresholds, a situation favorable for acceleration due to in-memory processing.

3. PROBLEM DEFINITION

In this section, we introduce the definitions of frequent pattern mining in a distributed computing environment, and we present the concept of Entropy and Quartiles.

3.1. FREQUENT PATTERNS

Consider an itemset with n different items, $I = \{x_1, x_2, x_3, ..., x_n\}$. A pattern or itemset is a subset of this itemset, represented as $X \subseteq I$. Since each transaction T is a subset of I and is uniquely recognized by its transaction ID (TID), a database DB is basically a collection of transactions. The amount of DB transactions that contain a pattern X, shown as support (X), indicates whether or not it is supported. The formula is calculated as follows [2]:

$$support\ (X) = \frac{frequency(X)}{N} \quad (1)$$

where $count(X)$ is the frequency of X in the database and N is the total number of transactions. If a pattern's support achieves or exceeds a minimal support level, shown by the symbol ξ, it appears to be frequent. Finding every pattern that frequently appears in a transaction database while staying within the specified support threshold ξ is the primary goal of frequent pattern mining. For many data analytics applications, this process is crucial, particularly when it comes to finding significant patterns and correlations in enormous quantities of data.

The two main approaches in frequent pattern mining are the pattern growth method and the candidate set generation method. The choice between these methods is primarily determined by the size and complexity of the dataset, even though each offers special advantages in certain circumstances.

- Candidate Set Generation Method [8]: This conventional approach is a systematic process that generates and examines candidate sets to identify recurring patterns in a dataset. One popular technique that applies this idea is Apriori. Its foundation is the idea that no pattern that is not prevalent in the database cannot also be frequent if it is formed from a pattern of length (k+1). All of a pattern's subsets must similarly satisfy the necessary frequency threshold in order for it to be considered frequent. Even though this method works well for identifying significant patterns, it can become computationally difficult, particularly when working with big datasets.
- Pattern Growth Method [7]: This method, in contrast to the candidate set generation method, uses previously found frequent patterns to avoid the requirement to build candidate sets. It concentrates on identifying local, frequent patterns that are gradually extended to produce more extensive global patterns rather than methodically integrating every potential item. Its computational efficiency is one of this method's main advantages. Processing huge datasets is rendered easier by reducing the requirement for multiple database scans by converting the database into a more compact and memory-efficient structure.

3.2. FREQUENT PATTERNS IN DISTRIBUTED DATABASES

There are n sites in a distributed system, which are designated $S_1, S_2, S_3, ..., S_n$. Within this system, the database DB is horizontally partitioned into n segments, denoted as $DB_1, DB_2, DB_3, ..., DB_n$, where each partition DB_i is allocated to a specific site S_i for processing (for i = 1, ..., n). To evaluate the occurrence of a pattern X, we define $support_i(X)$ as its local support count within, DB_i, while $support\ (X)$ represents its global support count across the entire distributed system. If a pattern X satisfies the minimal support criterion min_{sup}, which is established by the formula:

$$support\ (X) \geq min_{sup} \times |DB| \quad (2)$$

it can be considered globally frequent. In a distributed database setting, this condition is important since it assures that the pattern appears frequently across several partitions.

3.3. CONCEPT OF ENTROPY AND QUARTILES

In information theory, entropy is a measure of the uncertainty of a random variable. In this study, entropy is used to measure the distribution of items and enable the removal of branches in prefix trees. The conventional Shannon [19] formula for entropy is given by (Equation 3):

$$H(X) = -\sum_{i=1}^{n} p(X_i) log_2 p(X_i) \quad (3)$$

$P(X)$ is the probability that an item X is present in a transaction. Takes into account only the occurrence of an item. Here, we use its binary form [20], which also considers the non-occurrence; the formula can be expressed as (Equation 4):

$$H(X) = -P(X) log_2 p(X) - \big(1 - p(X)\big) log_2 \big(1 - p(X)\big) \quad (4)$$

$P(X)$ is the probability that a branch X is present in a transaction. $1-P(X)$ is the probability that X is absent. This formula measures the uncertainty or impurity related to the presence or absence of X. If the item or branch is very frequent or very rare, entropy is low (less uncertainty). If the item or branch appears about 50% of the time, entropy is high (maximum uncertainty).

Each branch is assessed using this measure to ascertain how informative it is. To filter out less relevant branches, we employ statistical quartile-based adaptive pruning. Level $Q1$ (25%): immediate elimination of branches whose entropy is lower than the first quartile ($Q1$). Median Level $Q2$ (50%): removal of the remaining branches whose entropy is lower than the median, but only partially. Level $Q3$ (75%): exclusion of branches with entropy values falling below the third quartile ($Q3$), thereby including only highly relevant and informative branches. The ranks are calculated according to the formula (Equation 5) [21]:

$$Q = q \times (N - 1) + 1 \quad (5)$$

Where q corresponds to 0.25 ($Q1$), 0.50 ($Q2$), or 0.75 ($Q3$), and N is the total number of values in the sample sorted in ascending order. When the position obtained is decimal, linear interpolation is applied in accordance with the method developed by Walpole *et al.* [22] as follows:

$$Q = V_{inf} + f \times (V_{sup} - V_{inf}) \quad (6)$$

In this case, V_{inf} is the value at rank i in the ranked list, and V_{sup} is the value at rank i+1. The letter f indicates the fractional part of the calculated rank. This approach provides a gradual reduction of branches, while still preserving those of high information content.

4. PROPOSED APPROACH

In this section, we present the AGFPM approach, developed according to a Master-Slave architecture adapted to distributed environments [32]. The process begins with the division of the transactional database into several fragments, each assigned to a slave site. Each Slave site then ensures autonomous processing of its portion of data by locally constructing a prefix tree (Local Prefix-Tree) in a parallel manner. A consolidation step then occurs to aggregate the local structures into a unified global tree (Global Prefix-Tree). Unlike classical methods that rely solely on absolute frequencies, this fusion is based on a binary entropy measure, allowing for the evaluation of the informational contribution of each subtree. The approach thus allows for the efficient extraction of global frequent patterns from this consolidated structure. The architectural representation of the entire process is presented in Fig.1.

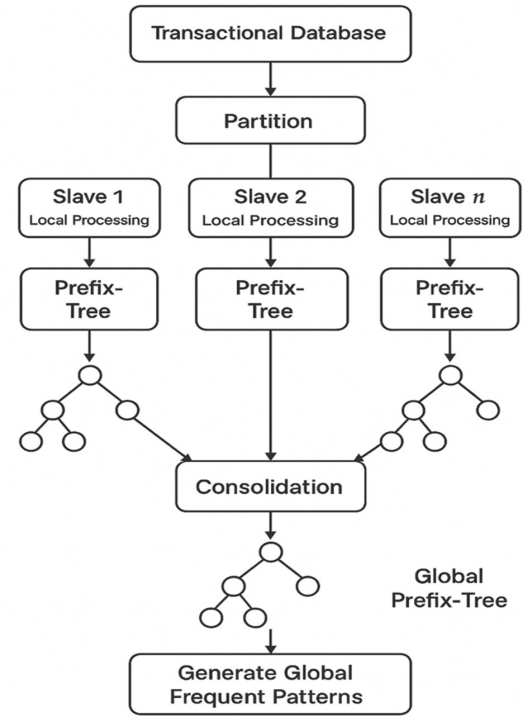


Fig. 1. General process of AGFPM algorithm

4.1. SCHEME OF COMMUNICATION

In the context of distributed algorithms, the Master-slave scheme constitutes a fundamental architecture that allows for the efficient coordination of operations between multiple processing nodes. This model relies on a clear separation of roles: the Master site assumes the responsibility of the overall orchestration of the process, while the slave sites locally execute the assigned tasks, such as data preprocessing, frequent itemset extraction, or association rule generation. This organization promotes a significant reduction in communication costs, a crucial aspect in distributed environments where bandwidth can be a limiting factor.

One of the major advantages of this scheme lies in its ability to minimize computation duplications, centralize aggregations, and promote a rapid convergence towards coherent global results, even when the data is massive and distributed heterogeneously. Moreover, the model facilitates the implementation of load balancing policies by dynamically distributing data partitions or subtasks according to the capacity or state of each Slave site. In the specific field of frequent pattern mining, this paradigm is commonly adopted for its ease of implementation, robustness, and compatibility with traditional distributed infrastructures (clusters, clouds).

Works such as those of Oliveira and Zaïane [33] or more recently Tseng *et al.* [34] have demonstrated that the Master-Slave scheme, combined with compact data structures like FP trees or prefix lattices, optimizes overall performance while ensuring low latency in the transfer of intermediate results.

Our algorithm relies on this scheme of communication for optimizing frequent pattern extraction. Each site constructs an LP-Tree locally, and the Master aggregates the supports and constructs a global GP-Tree. The use of entropy measures and adaptive thresholds (quartiles) allows for statistically significant pruning, ensuring global coherence, accuracy of extracted patterns, and minimization of exchanges between sites.

4.2. DEVELOPMENT OF PARALLEL LOCAL PREFIX-TREE

The Local LP-Tree, or Local Prefix Tree, is a parallel structure developed for each partition of the distributed database. It follows a tree-like structure modeled on the FP-Tree and has a root node along with multiple subtrees that denote prefix paths; each node contains a local support counter and a pointer to similar nodes. A Local Header Table is built to facilitate the subsequent extraction of conditional patterns. The process of construction involves three main steps. First, each slave site sorts its transactions according to a predefined lexicographic order, inserts the prefixes into an LP-Tree, and records the counters in the local header table.

Then, in the second phase, all the counters are sent to the Master site, which calculates the binary Shannon

entropy of each prefix. In order to estimate its informational value. The prefixes are then globally ranked by decreasing entropy, and this ranking is used to restructure the branches of the local trees, thereby ensuring global coherence between the sites. The LP-Trees are then reorganized according to this new order and their local header tables updated. The use of the order of global entropies in our solution guarantees a meaningful tree, homogeneous at each node, and optimum efficiency for the subsequent construction of a global prefix tree to accurately mine globally frequent patterns.

In the third step, all Slave sites compute the binary Shannon entropy of their whole branches (paths from the root to the node) to approximate the informative quality of the branches in each LP-tree. Most frequent branches carry a higher probability. Less frequent branches carry a lower probability, which is the reason for their elimination during pruning.

To conduct the multi-level pruning of the LP-tree branches by their entropies, we establish a dynamic adaptive threshold for the pruning process; this is determined using the quartiles of the entropies calculated. The values of branch entropies are sent to the Master site, where they are merged into one list [$H(B_1)$, $H(B_2)$, $H(B_n)$], in order to calculate three adaptive thresholds as quartiles ($Q1$-25%, $Q2$-50%, $Q3$-75%), we arrange the entropies in ascending order as follows:

$$H(B_1) \leq H(B_2) \leq ... \leq H(B_n) \quad (7)$$

We then use the formula of Cover and Thomas [24] cited in subsection 3 of section 3 to calculate $Q1$, $Q2$ (median), and $Q3$. The first quartile ($Q1$, 25%) allows the immediate removal of branches with very low entropy. The second quartile (Median, 50%) removes systematically branches that have medium entropy. The third quartile ($Q3$, 75%) is used as a last resort to keep only the most informative branches.

The setting of these three thresholds enables a systematic, incremental, and statistically valid pruning process by ensuring that only branches of high informational value (defined by high entropy, reflecting a high diversity of items) are retained. Therefore, branches with entropy levels below these three thresholds are eliminated, thus allowing the incremental elimination of unnecessary and less valuable branches at different levels. The first quartile ($Q1$) level involves the rapid elimination of branches with entropy values below this first quartile. Median Level ($Q2$) is the elimination of branches whose entropy levels fall below the median. Level $Q3$ is the stringent elimination of branches whose entropy is still lower than the third quartile ($Q3$), so that only the most informative and relevant branches are retained. This improved method supports the precise and efficient strengthening of the LP-tree to obtain an optimized LP-tree, avoiding overly aggressive removal of useful information and maintaining a balance between memory savings and preservation of significant patterns.

Finally, from the optimized LP-Trees, each site extracts a local database of conditional patterns (CL), containing for each prefix its ancestors and their associated supports. These databases contain the essential information and will act as a basis for the merging of patterns on a global scale in the subsequent steps. Algorithm 1 gives the phases involved in our LP-Parallel Tree construction.

Algorithm 1: Parallel Construction of LP-Tree

Input:

$BD \leftarrow$ Global transactional database

$N \leftarrow$ Number of distributed sites

$min_{sup} \leftarrow$ Minimum support threshold

Output:

LP-Tree-Optimized[i] ← Pruned local prefix trees for each site S_i

$CL[i] \leftarrow$ Local conditional pattern bases from each site

1. // *Step 1*: Horizontal partitioning of the database
2. Divide BD into N disjoint subsets $\{BD_1, BD_2, ..., BD_n\}$
3. Distribute BD_i to each corresponding site S_i
4. // *Step 2*: Local construction of *LP-Trees*
5. **for each** site S_i **do**
6. Build initial *LP-Tree* [i] from BD_i (prefix-based tree)
7. Construct *Local-Header-Table* [i] from *LP-Tree* [i]
8. Each node stores: prefix, local counter, and node links
9. Send local prefix counters to the Master site
10. **end for**
11. // *Step 3*: Global entropy computation
12. Master aggregates all local counters
13. **for each** prefix P_i **do**
14. $P(p_i) \leftarrow$ frequency $(P_i) / |BD|$
15. $H(p_i) \leftarrow -P(p_i) \cdot \log_2(P(P_i)) - (1 - P(p_i)) \cdot \log_2(1 - P(p_i))$
16. **end for**
17. Sort prefixes by decreasing entropy $H(P_i)$
18. Send the sorted prefix order to all sites
19. // *Step 4*: Tree reconstruction using global entropy order
20. **for each** site S_i **do**
21. Reorder transactions based on entropy order
22. Rebuild LP-Tree [i] accordingly
23. Rebuild Local-Header-Table [i] from updated LP-Tree [i]
24. **end for**
25. // Step 5: Compute branch entropies
26. **for each** site S_i **do**
27. **for each** branch B_j in *LP-Tree* [i] **do**
28. Compute entropy $H(B_j)$ using Shannon binary entropy

29. **end for**
30. Send all branch entropies to the master
31. **end for**
32. // *Step 6*: Global adaptive thresholds (quartiles)
33. Master aggregates all entropies and computes:
34. $Q_1 \leftarrow 25\%$; $Q_2 \leftarrow 50\%$; $Q_3 \leftarrow 75\%$
35. Send Q_1, Q_2, Q_3 to all sites
36. // *Step 7*: Multilevel pruning based on entropy
37. **for each** site S_i **do**
38. Remove branches with $H(B) <$ all thresholds (Q_1, Q_2, Q_3)
39. *LP-Tree-Optimized* [*i*] ← resulting pruned tree
40. **end for**
41. // *Step 8*: Extract local conditional patterns
42. **for each** *LP-Tree-Optimized* [*i*] **do**
43. Extract *CL* [*i*] ← Conditional pattern base for each prefix
44. **end for**

4.3. DEVELOPMENT OF GLOBAL PREFIX-TREE

In this section, we focus on the global knowledge from the local structures by constructing a global compact prefix arborescence from the pruned local LP-Trees. Our algorithm AGFPM is responsible for focusing this distributed information by building a global prefix tree known as GP-Tree. The Master site receives all the CL extracted by the local sites. For each prefix, the corresponding ancestors (and their counters) are merged. When several sites share the same prefix and ancestor associations, their counters are aggregated. This phase builds a unified structure that gathers all the local data in a condensed format. After the aggregation process is finished, the GP-Tree is developed by adding each Ancestor → Prefix path into a central tree while keeping the aggregated counters by means of a Global Header Table (GHT) that encompasses all of the counters for each prefix. The result is a compact, centralized arborescent structure that represents the global distribution of frequent patterns. This structure is the foundation of more advanced pruning and extraction processes that are part of the approach. In order to accommodate various analysis environments, the method permits the user to select one of two GP-Tree processing modes. In the first instance, pruning can optionally be used to cut down on the complexity of the tree, thereby reducing the amount of extracted frequent patterns and concentrating the results on the most important branches.

The method employed is as follows: Each distinct path in the GP-Tree (root to node) is analyzed with Shannon's binary entropy to quantify the amount of uncertainty or information associated with this path. The objective is to identify the most important branches, i.e., those that carry the highest amount of information regarding the overall frequent patterns. Following the computation of the entropies of all the branches, the Master site proceeds to define the distribution and calculate three adaptive thresholds in the form of quartiles: Q_1 (25%), Q_2 (50% or median), and Q_3 (75%). These thresholds enable multi-level pruning to be carried out. To be precise, any branch with an entropy value that falls below these three thresholds (Q_1, Q_2, and Q_3) is assumed to be of very limited informational value and is therefore pruned. This approach enables retaining only the most informative patterns while, concurrently, minimizing the size of the tree and eliminating redundancies. The GP-Tree thus produced is not only more concise and informative but also more appropriate for mining high-quality global conditional patterns and thus reducing redundancy and increasing interpretability of global frequent patterns. In another mode, the user can opt to keep the entire GP-Tree, hence enabling an exhaustive pattern extraction without utilizing rigorous filtering techniques. Such flexibility enables the adjustment of analytical granularity to suit each application's individual requirements, either to performance or to comprehensiveness. GP-Tree is instrumental in the derivation of global conditional models, which ultimately enables the determination of common global patterns. In either mode, the last GP-Tree is utilized as the foundation for the construction of global conditional models (CG), which results in the extraction of global frequent patterns. Algorithm 2 describes the steps of constructing the GP-tree.

Algorithm 2: Construction of GP-Tree

Input:

CL [1...*N*] ← Local conditional pattern bases from N distributed sites

Output:

GP-Tree ← Global Prefix Tree constructed from merged local

CLs

CG ← Global conditional patterns

1. Initialize an empty dictionary Aggregated-*CL* ← {}
2. // *Step 1*: Aggregate local conditional patterns
3. **for** $i \leftarrow 1$ to N **do**
4. for each (prefix P, list of ancestors $\{A_1, A_2, ..., A_k\}$) in *CL* [*i*] **do**
5. **if** $P \notin$ Aggregated-CL **then**
6. *Aggregated-CL* [*P*] ← {}
7. **for each** ancestor A_j in $\{A_1, ..., A_k\}$ **do**
8. **if** $A_j \in$ *Aggregated-CL* [*P*] **then**
9. *Aggregated-CL* [*P*][A_j] ← *Aggregated-CL* [*P*][A_j] + *count* (*CL* [*i*][*P*] [A_j])
10. **else**
11. *Aggregated-CL* [*P*][A_j] ← *count* (*CL* [*i*] [*P*] [A_j])

12. **end for**
13. **end for**
14. **end for**
15. // *Step 2*: Construct the Global Prefix Tree (GP-Tree)
16. Initialize GP-Tree as an empty prefix tree with a root node
17. **for each** prefix P in $Aggregated\text{-}CL$ **do**
18. **for each** ancestor A_j in $Aggregated\text{-}CL$ [P] **do**
19. Insert the path $A_j \rightarrow P$ into GP-Tree
20. Update the node counters with $Aggregated\text{-}CL$ [P][A_j]
21. Construct GHT from GP-Tree // Global Header Table
22. **end for**
23. **end for**
24. // *Step 3*: Extract final global conditional patterns
25. CG ← Extract global conditional patterns from GP-Tree
26. **return** GP-Tree

4.4. GLOBAL FREQUENT PATTERNS MINING

In our approach, global conditional patterns (CG) are used (CG) to build GP subtrees (sub-GP-Trees) at various levels of iterations as K. Each sub-GP-Tree categorizes systematically frequent patterns at a specific level of granularity, thereby enhancing the detection of more complicated combinations, e.g., pairs or triplets. Unlike an FP-tree that explores elements from bottom to top, the sub-GP-Tree is traversed from top to bottom in the Global Header Table. This top-down exploration method allows the efficient derivation of all frequent global element sets related to X.

The procedure is iterative: At level $K = 1$, for each prefix of size (m=1) obtained from the global conditional patterns (CG) of GP-tree, a sub-GP-Tree is produced. For each frequent global pattern of size ($m = 2$) obtained from the initial level patterns, a sub-GP-Tree is built at level $K = 2$. It is done by building the sub-GP-Tree based on the intersection of the paths for subsets of global frequent patterns. The mining of common paths among subsets of X, each containing (m−1) elements and sharing common prefixes of length (m−2) at the start of the frequent global patterns X, constitutes the next step of the procedure at level K≥3. This procedure continues until no new sub-GP-Tree can be constructed, i.e., no further frequent global pattern can be identified. Likewise, the sub-GP-Tree of XY cannot be constructed if X is part of a global frequent pattern XY and does not have any ancestors, and therefore, prevents the construction of additional global frequent patterns. Algorithm 3 describes the procedure for obtaining global frequent patterns.

Algorithm 3: Extract Global Frequent Patterns

Input:

$GP\text{-}Tree$ ← Construction of $GP\text{-}Tree$ (from Algorithm 2, global header table implicitly built)

Output:

$G\text{-}Patterns$ ← Complete set of global frequent patterns

1. Initialize $G\text{-}Patterns \leftarrow \emptyset$ // Final collection of global frequent patterns
2. // *Recursive pattern mining starting at level* $K = 1$
3. **for** $K \leftarrow 1$ to ∞ do
4. New-Patterns ← ∅ // Temporary patterns discovered at level K
5. **if K** = 1 then
6. // *First level:* explore directly from GP-Tree structure
7. **for** each global frequent pattern X from $GP\text{-}Tree$ **do**
8. Build $Sub\text{-}GP\text{-}Tree$ [X] from conditional paths including X
9. **if** $Sub\text{-}GP\text{-}Tree$ [X] ≠ ∅ **then**
10. Extract frequent patterns from $Sub\text{-}GP\text{-}Tree$ [X]
11. Add them to New-Patterns and to G-Patterns
12. **end if**
13. **end for**
14. **else**
15. // *Next levels*: *explore from each global header table GHT of previous subtrees*
16. **for each** global frequent pattern X derived from $Sub\text{-}GP\text{-}Tree$ [X]. GHT **do**
17. Build $Sub\text{-}GP\text{-}Tree$ [X] from conditional paths including X
18. **if** $Sub\text{-}GP\text{-}Tree$ [X] ≠ ∅ **then**
19. Extract frequent patterns from $Sub\text{-}GP\text{-}Tree$ [X]
20. Add them to New-Patterns and to G-Patterns
21. **end if**
22. **end for**
23. **end if**
24. **if** New-Patterns = ∅ **then**
25. break
26. **end if**
27. // *Update header table of each* $Sub\text{-}GP\text{-}Tree$ [X] *for the next level*
28. Update GHT for each $Sub\text{-}GP\text{-}Tree$ [X] // Global Header Table: GHT
29. **end for**
30. **return** G-Patterns

Our proposed AGFPM algorithm is based on a succession of complementary steps, each one of which is essential for building the data structures required for the final extraction. The method permits concentrating computational resources on the most representative patterns using the techniques of the distributed organization of processing, focused use of entropy to prioritize information, and application of intelligent pruning. Both structural and informational, this optimization offers a strong basis for proceeding to the study of the results. It enables a valid assessment of the relevance of the identified patterns as well as a clear observation of the performance of the suggested approach in several experimental settings.

5. RESULTS AND DISCUSSION

We carried out comprehensive experiments on two types of datasets, as indicated in Table 1, with distinct properties in order to assess the performance of our AGFPM. The synthetic dataset T10I4D100K was generated by the IBM Almaden Quest group and obtained from [23]. The Kosarak dataset was used in its transaction format as distributed by SPMF [24]. The Chess dataset, derived from the UCI Machine Learning Repository, was used in its transaction-encoded form as distributed by SPMF [25]. We compared AGFPM against well-known methods such as PFP-Tree and CD. The tests were conducted on a PC running Windows 11 with an Intel® CoreTM i7-10875H CPU running at 2.80 GHz and 16 GB of RAM. The datasets were dispersed among 3, 5, and 7 Slave sites in order to evaluate scalability and efficiency. The Java program was generated with the NetBeans IDE. MPJ Express, a Java-based message passing framework created especially for running parallel applications on multicore machines, facilitates communication between sites.

Table 1. Dataset Characteristic

Dataset	Transaction	Items	Max TL (Maximum Tree Length)	Avg TL (Average Tree Length)
T10I4D100K	100000	870	29	10.1
Kosarak	990002	41270	2498	8.10
Chess	3196	75	38	36

5.1. ANALYSIS OF PERFORMANCES

The efficiency and performance of the PFP-Tree method, CD algorithm, and proposed AGFPM technique vary considerably in the case of the T10I4D100K, Kosarak, and Chess datasets. Fig. 2 presents a detailed comparison of the execution times of the AGFPM algorithm applied to the T10I4D100K dataset, based on different values of the minimum support threshold (MinSupp).

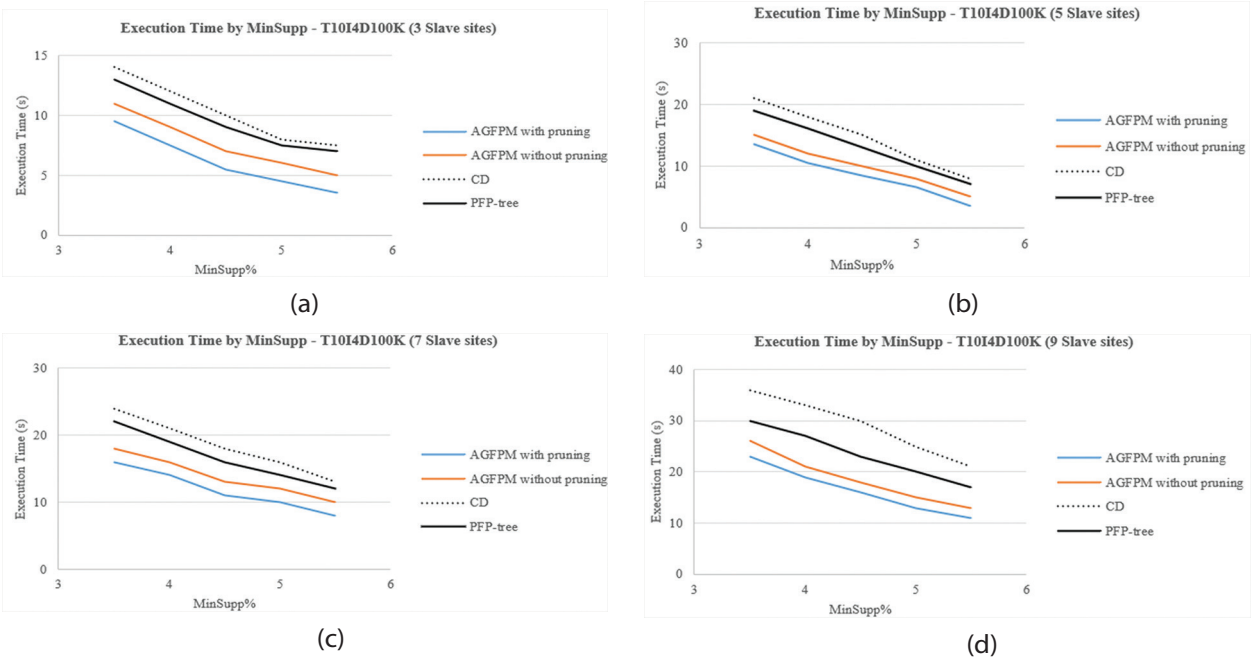


Fig. 2. The excution time of T10I4D100K with (a) 3 numbers of Slave sites, (b) 5 numbers of Slave sites, (c) 7 numbers of Slave sites, (d) 9 numbers of Slave sites

The study covers three distributed configurations, involving 3, 5, 7, and 9 slave sites respectively, in order to evaluate the impact of the degree of parallelism on performance. The displayed curves allow for a comparative analysis of the behavior of four algorithms: AGFPM with pruning, AGFPM without pruning, CD (Count Distribution), and PFP-tree (Parallel FP-tree), for MinSupp values ranging from 3.5% to 5.5%. This visualization provides a detailed analysis of the gains achieved by AGFPM compared to classical methods, while highlighting the influence of the number of slave nodes on execution speed and the processing capacity of the different models.

For a configuration with 3-slaves, the AGFPM algorithm with pruning shows the best results, with execution time decreasing from 9.5 s to 3.5 s. This improvement demonstrates the effectiveness of the entropy-based pruning mechanism, which reduces the complexity of the global tree by removing uninformative branches. In comparison, the version without pruning is slightly slower (between 11 s and 5 s), which confirms the value of adaptive filtering. The reference approaches are less effective: CD varies between 14 s and 7.5 s, while PFP-tree drops from 13 s to 7 s, revealing their limitations in contexts where parallelism is moderate.

When the number of slave sites is increased to 5, performance improves significantly. AGFPM with pruning reaches 13.5 s at MinSupp = 3.5%, and drops to 3.5 s at MinSupp = 5.5%. The gain becomes even more notable compared to other methods: CD remains above 21 seconds, and PFP-tree around 19 seconds at the minimal MinSupp. The stability of the differential between the two versions of AGFPM confirms that pruning maintains its effectiveness, regardless of the level of parallelism. This behavior validates the adaptability of the model to a medium-sized distributed architecture.

In the 7-slave configuration, although execution times continue to decrease (for example, AGFPM with pruning goes from 16 s to 8 s), the marginal gain diminishes. This is due to the increase in synchronization costs, particularly in the global aggregation phase. The CD algorithm caps at 24 seconds for the lowest threshold, while the PFP-tree remains at 22 seconds. These results show that excessive parallelism can sometimes induce an overhead that reduces the overall efficiency of the system, especially when the size of the partitions becomes too fine relative to the communication load between the master and the slaves.

Under the 9-slave configuration, the ordering is stable, and runtimes shrink as MinSupp increases. At 3.5%, AGFPM with pruning is 23s, clearly ahead of PFP-tree (30s) and CD (36s); by 5.0%, it falls to 13s while competitors remain higher. This pattern reflects a design that aligns local structures before aggregation and filters low-yield branches early, reducing both search space and cross-site reconciliation advantages that become more salient as coordination costs rise at this degree of parallelism.

Across all tested configurations, AGFPM with pruning stands out clearly, combining speed, scalability, and relevance of the extracted patterns. The use of pruning based on entropy quartiles allows for the dynamic regulation of the growth of the global tree, which lightens the processing without compromising the quality of the results. These observations highlight the relevance of the proposed approach in distributed environments, particularly when the goal is to efficiently process large volumes of data with low frequency thresholds.

Fig. 3 illustrates the evolution of the execution time of the various distributed algorithms applied to the Kosarak dataset, known for its high density.

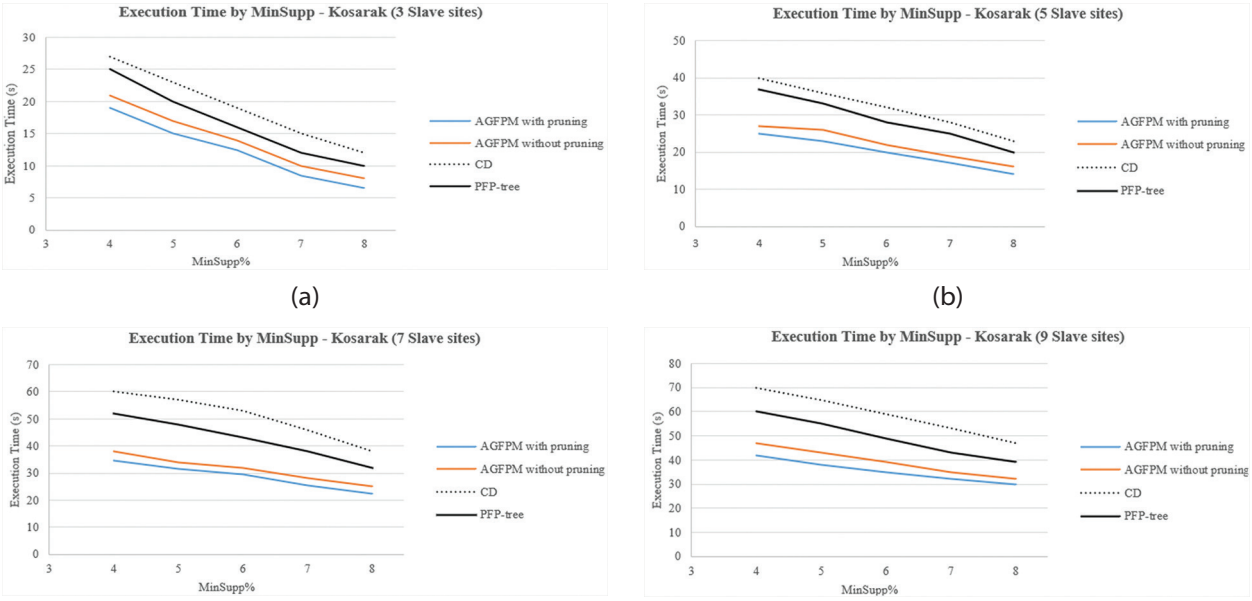


Fig. 3. The execution time of Kosarak with (a) 3 numbers of Slave sites, (b) 5 numbers of Slave sites, (c) 7 numbers of Slave sites, (d) 9 numbers of Slave sites

The results are analyzed for five increasing values of the minimum support threshold (from 4% to 8%) and at three levels of parallelism, involving 3, 5, and 7 slave sites. The performances of five algorithms are compared: AGFPM with pruning, AGFPM without pruning, CD, and PFP-tree.

In the configuration with 3 slave sites, the AGFPM algorithm with pruning stands out for its speed, with execution time decreasing from 19 seconds at MinSupp = 4% to 6.5 seconds at 8%. This efficiency is explained by the impact of the pruning strategy, which significantly reduc-

es the volume of patterns to be aggregated. The version without pruning is slower, reaching 21 seconds at 4% and 8 seconds at 8%, which confirms the value of informational filtering. The CD and PFP-tree algorithms show significantly higher execution times: for example, CD starts at 27 seconds at 4% and remains at 12 seconds at 8%, revealing a low adaptability to transaction density.

With 5 slaves, the gap between the algorithms becomes more pronounced. AGFPM with pruning remains the most efficient, with a time of 25 seconds at 4%, dropping to 14 seconds at 8%. In comparison, CD requires 40 seconds at 4% and 28 seconds at 8%, proving its low scalability in high-density environments. PFP-tree, although improved by parallelism, remains above 20 seconds even for high MinSupp values. The stability of AGFPM's performance with pruning confirms its ability to exploit parallelism while maintaining low computational overhead.

When the approach is deployed on 7 slaves, the observed trends are confirmed. The execution times remain controlled for AGFPM with pruning, which goes from 34.5 seconds at MinSupp = 4% to 22.5 seconds at 8%. The competing algorithms, on the other hand, show obvious limitations. CD caps at 60 seconds at 4% and struggles to go below 38 seconds, despite the increase in the number of nodes. PFP-tree follows a similar trend (from 52s to 32s).These results highlight the structural limitations of these methods when they have to manage a high density in a massively distributed context.

Across 9 slave sites, the ordering remains consistent, and runtimes decrease steadily as MinSupp rises. At 4%, AGFPM with pruning records 42s, clearly below PFP tree (60s) and CD (70s); by 8%, it reaches 30s while competing methods remain higher (PFP tree 39s, CD 47 s). This trajectory indicates that the pruned variant sustains a stable margin across the range, with the gap most visible at lower supports where the search space and coordination overheads are largest. The advantage arises from a design that harmonizes local structures before aggregation and suppresses low-yield branches early, thereby reducing both conditional growth and cross-site reconciliation effects that are particularly beneficial on a large, sparse workload such as Kosarak under high parallelism.

In summary, the results of Fig. 3 demonstrate the consistent superiority of the AGFPM approach with pruning, which combines algorithmic efficiency, informational compression, and intelligent exploitation of parallelism. This method adapts better than its competitors to data density and support variations, while maintaining robust scalability in the face of an increasing number of slave sites. It thus constitutes a relevant solution for the mining of global frequent patterns in complex distributed environments.

Fig. 4 depicts how execution time varies with MinSupp across four algorithms: AGFPM with pruning, AGFPM without pruning, CD (Count Distribution), and PFP-tree on the Chess dataset as the number of slave sites increases. Across all subfigures, higher MinSupp thresholds shorten runtime, while the performance ranking remains consistent: AGFPM with pruning is fastest, followed by its no-pruning variant, with PFP-tree and CD trailing gaps widening at low MinSupp where dense co-occurrences inflate intermediate structures.

Under the 3-slave setup, method separations are most visible at low MinSupp and narrow as the threshold rises. At MinSupp=0.12%, AGFPM with pruning is about 0.085s and keeps a lead that persists through mid supports (≈0.20).

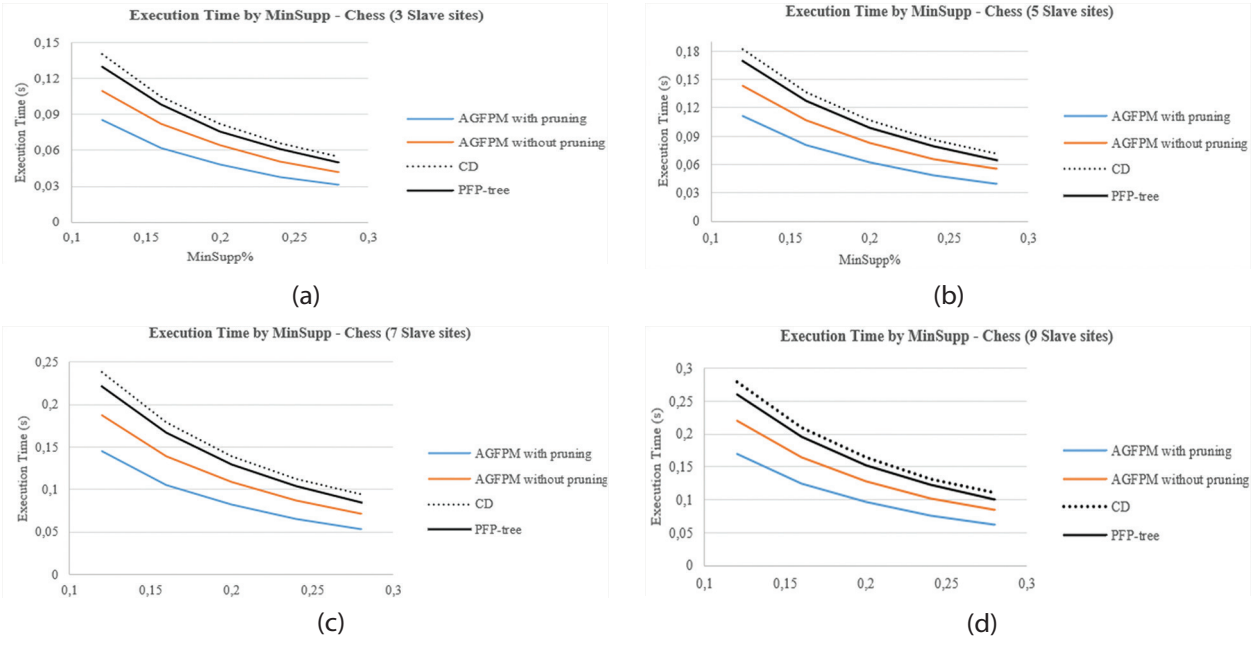


Fig. 4. The execution time of Chess with (a) 3 numbers of Slave sites, (b) 5 numbers of Slave sites, (c) 7 numbers of Slave sites, (d) 9 numbers of Slave sites

The ordering remains stable, pruned variant first, then no-pruning, with PFP-tree and CD behind. This reflects an information-guided item order that reduces cross-site mismatch and multi-level pruning that curbs conditional growth, lowering counting and merge costs, while PFP-tree and CD incur, respectively, higher merging and candidate/synchronization overheads at the smallest supports.

With 5 slave sites, the ranking remains stable and the gaps widen at low MinSupp, indicating rising structural and coordination costs as parallelism increases. At MinSupp=0.12%, AGFPM with pruning is clearly ahead (≈0.11s) and maintains a sizeable margin over its no-pruning variant and both PFP-tree and CD. The advantage persists at mid supports (e.g., around 0.20%), where the pruned version remains the fastest and the no-pruning variant stays competitive, while PFP-tree and CD trail. This efficiency is explained by information-guided item ordering that aligns local structures before aggregation and multi-level pruning that restrains conditional growth; together, they reduce merging and synchronization overheads that otherwise escalate for PFP-tree (structure merging) and CD (candidate generation and repeated coordination).

Under the 7-slave setting, separations are largest at low MinSupp and narrow as the threshold increases. At MinSupp=0.12%, AGFPM with pruning is around 0.145s and clearly ahead of its no-pruning variant, PFP-tree, and CD; the ordering remains unchanged at mid supports (e.g., 0.20%). This advantage stems from information-guided item ordering that aligns local structures before combination and multi-level pruning that limits conditional growth. By contrast, PFP-tree's merging cost and CD's candidate/synchronization overheads scale poorly at low supports and higher parallelism, keeping their curves steeper in this regime.

With 9 slave sites, the relative ranking is unchanged, and the advantage of AGFPM with pruning is most evident under low MinSupp. Communication and coordination increasingly dominate overall cost at this scale; by transmitting compact metadata and imposing a consistent global item order, AGFPM curtails the information that must be reconciled across sites. The multi-level, quartile-guided filtering further limits conditional expansions, sustaining favorable runtime trends as parallelism grows.

On dense transactional data such as Chess, especially at low MinSupp and higher degrees of parallelism, AGFPM with pruning provides the most stable and efficient behavior. Its combination of information-guided ordering, compact exchanges, and adaptive filtering translates into reduced merging and synchronization costs, more predictable scaling from 3 to 9 slave sites, and a consistently lower execution-time profile than PFP-tree and CD.

The volume of motifs can be analyzed in Table 2, Table 3, and Table 4 according to five different support thresholds and for each algorithm (CD, PFP-Tree, AGFPM with and without pruning), for the T10I4D100K, Kosarak, and Chess datasets.

Table 2. Number of Frequent Patterns Mined for Kosarak

Algorithm	4% Min Supp	5% Min Supp	6% Min Supp	7% Min Supp	8% Min Supp
AGFPM with pruning	5228	5107	4485	4110	3698
AGFPM without pruning	5311	5283	4555	4275	3475
CD	5728	5314	5151	5018	4800
PFP-Tree	5546	5123	5000	4767	4105

Table 3. Number of Frequent Patterns Mined for T10I4D100K

Algorithm	3.5% Min Supp	4% Min Supp	4.5% Min Supp	5% Min Supp	5.5% Min Supp
AGFPM with pruning	2373	2254	2119	1822	1476
AGFPM without pruning	2420	2339	2275	2120	1653
CD	3532	3258	3100	2508	2003
PFP-Tree	2891	2670	2450	2254	1854

Table 4. Number of Frequent Patterns Mined for Chess

Algorithm	0.12% Min Supp	0.16% Min Supp	0.2% Min Supp	0.24% Min Supp	0.28% Min Supp
AGFPM with pruning	5210	4620	3980	3410	2890
AGFPM without pruning	5480	4840	4210	3620	3050
CD	5728	5110	4450	3880	3280
PFP-Tree	5610	4960	4320	3730	3160

Table 2 shows that AGFPM with pruning effectively limits the number of patterns extracted on a dense dataset like Kosarak. Compared to other algorithms, particularly CD and PFP-tree, it produces a more compact set of patterns, which facilitates interpretation and reduces redundancy. Entropy-based pruning thus allows targeting the most relevant patterns while maintaining good data coverage.

On the T10I4D100K dataset, Table 3 confirms the previously observed trend: AGFPM with pruning extracts fewer patterns than the other methods, particularly at low support thresholds. This ability to contain the size of the results is valuable for optimizing analysis, especially in distributed environments. In comparison, CD and PFP-tree generate a higher volume of rules, which can lead to an overload during post-processing.

Table 4 summarizes the number of frequent patterns identified in the Chess dataset across a range of minimum support thresholds. As the threshold increases, the volume of discovered patterns declines, consistent with a stricter inclusion criterion. Across methods, AGFPM with pruning consistently yields the most compact pattern sets, reflecting the effectiveness of early elimination of redundant or low-utility branches. The variant without pruning systematically retains more patterns,

illustrating the cost of deferring filtering. Classical baselines such as CD and PFP-tree produce larger outputs at all thresholds, indicating greater retention of redundant structures and, consequently, higher computational and storage overheads.

Overall, the three tables consistently demonstrate the superiority of AGFPM with pruning in controlling the volume of extracted frequent patterns. Across the evaluated datasets—Kosarak (large and sparse), T10I4D100K (lower density), and Chess (dense)—AGFPM produces markedly more compact pattern sets than classical baselines such as CD and PFP-tree. This conciseness does not come at the expense of result quality: the pruning strategy is guided by an entropy-based criterion that preserves the most informative branches of the global tree, ensuring that salient patterns are retained while redundant structures are discarded.

Compared with the non-pruned variant, the integrated filtering mechanism prevents the extraction of irrelevant or redundant patterns, an advantage that is especially valuable in distributed mining, where unnecessary candidates amplify communication, memory, and merge overheads. By contrast, traditional baselines, while sometimes competitive on specific datasets, tend to generate voluminous pattern sets that complicate interpretation and increase post-processing effort.

In summary, the quantitative results indicate that information-based adaptive pruning improves both computational efficiency and the clarity of the outcomes, yielding more compact and actionable pattern sets. This makes the approach well-suited to large transactional databases in distributed environments. Crucially, it reduces complexity without materially sacrificing informative content, which is a meaningful benefit for decision-support pipelines and time-sensitive analytics.

5.2. SCALABILITY ANALYSIS

In this section, we assess the scalability of AGFPM in a distributed setting by measuring execution time as the number of computing nodes varies (3, 5, 7, 9). The minimum support is set to 4% for T10I4D100K and Kosarak and 0.20% for Chess. The datasets span contrasting regimes: T10I4D100K (lower density), Kosarak (larger and sparser), and Chess (dense). We compare AGFPM with and without pruning against two classical baselines: CD (Count Distribution) and PFP-tree (Parallel FP-tree). The figures depict how increasing parallelism affects runtime and reveal each method's capacity to control coordination, merging, and candidate-handling overheads, thereby indicating their efficiency in a distributed environment.

Fig. 5 (a): On the T10I4D100K dataset, execution time grows steadily as the number of nodes increases from 3 to 9. The proposed AGFPM method, enhanced with multi-level pruning, consistently achieves the best performance, requiring only 7.5 seconds with 3 nodes and scaling efficiently to 19 seconds with 9 nodes. In contrast, the same method without pruning starts at 9 seconds and reaches 21 seconds, demonstrating the overhead caused by unfiltered, redundant branches. The performance gap widens further when compared to classical approaches: CD increases from 12 to 33 seconds, and PFP-tree from 11 to 27 seconds over the same range. These results underscore the effectiveness of adaptive pruning in reducing both computational load and communication volume, particularly as the system expands to larger cluster configurations. By eliminating non-promising paths early and limiting data exchange, AGFPM maintains efficiency even under increased parallelism.

Fig. 5 (b): The high-density and high-volume Kosarak dataset enhances performance gaps between the algorithms. The AGFPM with pruning continues to provide the best performance, going from 19 s with 3 nodes to 42s with 9 nodes. Without pruning, the durations rise more abruptly, from 21s to up to 47s, validating the overhead imposed by processing unfiltered data. The traditional approaches have much higher costs: CD rises sharply from 27s at 3 nodes to 70s at 9 nodes, and PFP-tree goes from 25s to 60s. The increasing performance gap between AGFPM with pruning and the others demonstrates the scalability of the pruning approach, which becomes ever more essential when the number of nodes increases and inter-node communications accelerate.

Fig. 5 (c): On the dense Chess dataset (MinSupp=0.20%), absolute runtimes are sub-second, yet the performance ordering is clear and consistent across node counts. AGFPM with pruning remains the fastest, roughly doubling from a few hundredths of a second at 3 nodes to under a tenth at 9 nodes, while the unpruned variant, PFP-tree, and CD occupy progressively higher bands at each configuration. Although the gaps are measured in milliseconds, they are systematic, indicating that pruning and a consistent item order curb conditional expansions and reduce reconciliation overhead. This stability shows that the approach is effective not only under heavy load but also in lighter regimes, where disciplined pruning still yields measurable, reproducible gains.

Overall, the findings indicate that AGFPM with pruning is particularly well suited to distributed settings. Curbing unnecessary expansions and harmonizing structures across nodes, it reduces computational load while scaling effectively with additional resources, which helps sustain low and stable runtimes even on large or highly correlated datasets.

In contrast to classical baselines that are prone to communication overhead and structural redundancy as parallelism increases, AGFPM combines efficiency, robustness, and adaptability. These properties make it a strong candidate for large-scale frequent pattern mining in modern distributed environments.

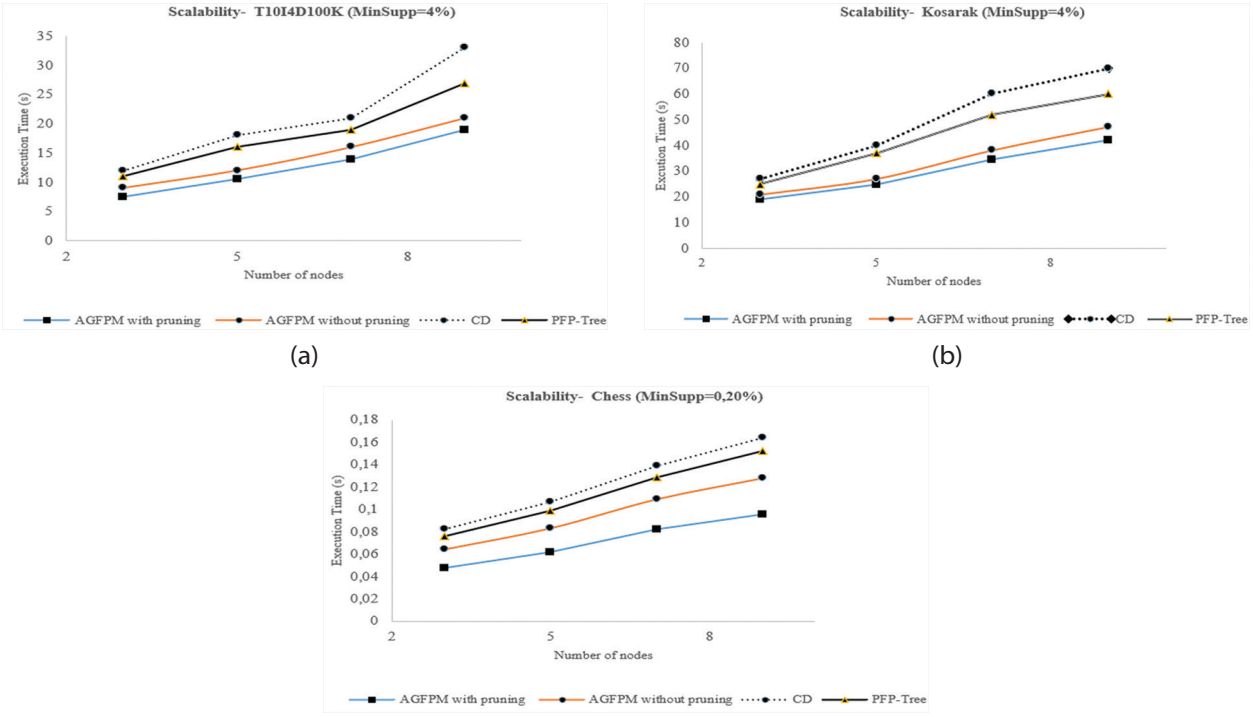


Fig. 5. Scalability of AGFPM by various number of nodes for (a) T10I4D100K, (b) Kosarak with MinSupp = 4% and (c) Chess with MinSupp = 0.20%

5.3. COMPARISON WITH RELATED METHOD ANALYSIS

We conduct a structured comparison of the main studied frequent pattern extraction methods, paralleling their performances according to several relevant evaluation criteria. The objective is to identify the strengths and limitations of each approach within a distributed and performance-oriented framework.

Table 5 summarizes this analysis based on four fundamental axes: the adopted architecture (centralized or distributed), the execution time observed during the experiments, the quality of the extracted patterns (in terms of relevance and reduction of redundancy), as well as scalability, measured through the algorithms' ability to maintain good performance when increasing the number of computing nodes.

Table 5. Detailed Comparison of Frequent Pattern Mining Methods

Criterion	CD	PFP-Tree	AGFPM without pruning	AGFPM with pruning
Architecture	Peer-to-peer or centralized, with inter-node communication repeated.	Mainly decentralized, yet synchronization is costly at the stage of merging.	Master-Slave structure; one master collects full, unfiltered local trees.	Master-Slave with metadata fusion and pruning optimized; reduces communication and central load.
Execution Time	High runtime, particularly at low support, because of repeated scanning of the database.	Moderate, but the growth of trees is expensive for large datasets.	Faster than CD/PFP, but time increases with dataset size and complexity.	Fastest overall, due to early pruning of low-entropy branches.
Pattern Quality	Correct patterns, but may not be redundant or noisy.	Robust patterns, sensitive to support threshold.	Very high-accurate pattern coverage, but possibly with irrelevant or marginal patterns.	Maintains accurate pattern quality and removes redundant or low-interest patterns.
Scalability (with more nodes)	Limited scalability; synchronization overhead rises with the number of nodes.	Moderate scalability; local tree merging is dependent.	Good scalability, but it can be affected by memory use.	Very good scalability; pruning decreases global tree size and inter-node communication.

In light of this comparison, the AGFPM algorithm, particularly in its version incorporating an adaptive pruning mechanism, confirms its superiority. It stands out for its execution speed, the informative quality of the results produced, and its ability to effectively adapt to large-scale distributed environments, making it a robust and efficient alternative to classical methods.

5.4. DISCUSSION

The experiments conducted on the T10I4D100K, Kosarak, and Chess datasets in a distributed setting (3, 5, 7, and 9 nodes) provide a comprehensive assessment of the AGFPM algorithm in its pruned and unpruned variants. The evaluation focuses on two dimensions: execution time and the number of extracted patterns, each

examined under varying minimum support thresholds. This design enables a balanced analysis of scalability and result compactness across heterogeneous data regimes.

The results clearly demonstrate that the integration of a pruning mechanism based on binary entropy and quartiles offers a significant advantage to AGFPM. By filtering out less informative branches before the global merging phase, the algorithm significantly reduces the size of the global tree to be processed. This reduction results in a significant decrease in execution times, particularly noticeable when the number of nodes increases or when the support is low. In some cases, the observed processing time is reduced by 30 to 40% compared to classical approaches such as CD or PFP-tree.

From a quantitative perspective, pattern extraction follows an expected logic: lower support thresholds generate more frequent patterns. Without a filtering mechanism, the unpruned version of AGFPM systematically produces the largest number of patterns, which can burden the analysis process. On the other hand, the pruned version retains most of the informative patterns while eliminating those with low added value. For example, for a support of 4%, the difference in the number of patterns between the two AGFPM variants is relatively small, while the gain in computation time is significant.

The CD and PFP-tree algorithms, although effective in certain contexts, show their limitations here: their lack of adaptability to low thresholds and their inability to efficiently handle data density result in longer response times and more cumbersome pattern sets.

In terms of efficiency, scalability, and overall performance on three datasets, the AGFPM algorithm particularly in its pruning version, stands out as the most robust solution. Its Master-slave scheme, combined with an optimized bidirectional communication strategy, reduces synchronization overhead and avoids redundant calculations. Thanks to this design, the approach manages to extract globally frequent patterns in a targeted manner, while maintaining short execution times and excellent scalability. These results confirm that the AGFPM algorithm represents a significant advancement for the efficient and parallel exploration of frequent patterns in large-scale distributed environments.

6. CONCLUSION

In large-scale data mining, extracting frequent patterns from distributed databases is constrained by communication overhead and input/output coordination. To address these limitations, we propose AGFPM (Adaptive Global Frequent Pattern Mining), a master–slave distributed algorithm that integrates two complementary structures: the LP-Tree for local discovery and the GP-Tree for global consolidation. This design reduces inter-site communication and enables efficient detection of frequent patterns at scale. The GP-Tree supports systematic mining of globally frequent patterns by iteratively building conditional sub-GP-Trees across hierarchical levels. This recursive process examines candidate itemsets in a targeted manner, allowing the method to adapt to data and pattern complexity without incurring unnecessary computational or memory costs. To further enhance relevance and control the complexity of both LP-Tree and GP-Tree, AGFPM employs an adaptive pruning strategy that couples binary entropy with quartile-based thresholds. This mechanism automatically filters low-informative branches while preserving highly correlated patterns, improving the quality and compactness of results in distributed environments.

Empirical results show that AGFPM, augmented with the pruning mechanism, delivers superior scalability and runtime performance relative to benchmark methods, while preserving high fidelity in identifying globally frequent patterns. Future work will extend this framework to the discovery of distributed association rules within a fully decentralized architecture. The aim is to refine the methodology and design an optimized algorithm that builds on AGFPM's foundations to further improve performance in distributed computing settings.

7. REFERENCES

[1] H. Kargupta, C. Kamath, P. Chan, "Distributed and Parallel Data Mining: Emergence, Growth, and Future Directions", Advances in Distributed and Parallel Knowledge Discovery, AAAI/MIT Press, 2000, pp. 409-416.

[2] R. Agrawal, T. Imieliński, A. Swami, "Mining Association Rules Between Sets of Items in Large Databases", Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, USA, 25-28 May 1993, pp. 207-216.

[3] P.-N. Tan, M. Steinbach, V. Kumar, "Association Analysis: Basic Concepts and Algorithms", Introduction to Data Mining, Pearson Addison Wesley, 2005, pp. 327-386.

[4] S. K. Tanbeer, C. F. Ahmed, B.-S. Jeong, Y.-K. Lee, "Efficient Single-Pass Frequent Pattern Mining Using a Prefix-Tree", Information Sciences, Vol. 179, No. 5, 2009, pp. 559-583.

[5] H. Huang, X. Wu, R. Relue, "Association Analysis with One Scan of Databases", Proceedings of the IEEE International Conference on Data Mining, Maebashi, Japan, 9-12 December 2002, pp. 629-632.

[6] G. Grahne, J. Zhu, "Fast Algorithms for Frequent Itemset Mining Using FP-Trees", IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 10, 2005, pp. 1347-1362.

[7] J. Han, J. Pei, Y. Yin, "Mining Frequent Patterns Without Candidate Generation", Proceedings of the ACM SIGMOD International Conference on Management of Data, Dallas, TX, USA, 16-18 May 2000, pp. 1-12.

[8] R. Agrawal, R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases", Proceedings of the 20th

International Conference on Very Large Data Bases, Santiago de Chile, Chile, 12-15 September 1994, pp. 487-499.

[9] R. Agrawal, J. C. Shafer, "Parallel Mining of Association Rules", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, 1996, pp. 962-969.

[10] D. W. Cheung, V. T. Ng, A. W. Fu, Y. Fu, "Efficient Mining of Association Rules in Distributed Databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, 1996, pp. 911-922.

[11] D. W. Cheung, J. Han, V. T. Ng, A. W. Fu, Y. Fu, "A Fast Distributed Algorithm for Mining Association Rules", Proceedings of the 4th International Conference on Parallel and Distributed Information Systems, Miami Beach, FL, USA, 18-20 December 1996, pp. 31-42.

[12] M. Z. Ashrafi, D. Taniar, K. Smith, "ODAM: An Optimized Distributed Association Rule Mining Algorithm", IEEE Distributed Systems Online, Vol. 5, No. 3, 2004, pp. 1-18.

[13] A. Schuster, R. Wolff, "Communication-Efficient Distributed Mining of Association Rules", ACM SIGMOD Record, Vol. 30, No. 2, 2001, pp. 473-484.

[14] E.-H. Han, G. Karypis, V. Kumar, "Scalable Parallel Data Mining for Association Rules", ACM SIGMOD Record, Vol. 26, No. 2, 1997, pp. 277-288.

[15] T. Shintani, M. Kitsuregawa, "Hash Based Parallel Algorithms for Mining Association Rules", Proceedings of the 4th International Conference on Parallel and Distributed Information Systems, Miami Beach, FL, USA, 18-20 December 1996, pp. 19-30.

[16] L. Harada, N. Akaboshi, K. Ogihara, R. Take, "Dynamic Skew Handling in Parallel Mining of Association Rules", Proceedings of the 7th ACM International Conference on Information and Knowledge Management, Bethesda, MD, USA, 3-7 November 1998, pp. 76-85.

[17] A. Javed, A. Khokhar, "Frequent Pattern Mining on Message Passing Multiprocessor Systems", Distributed and Parallel Databases, Vol. 16, 2004, pp. 321-334.

[18] O. R. Zaïane, M. El-Hajj, P. Lu, "Fast Parallel Association Rule Mining Without Candidacy Generation", Proceedings of the IEEE International Conference on Data Mining, San Jose, CA, USA, 29 November - 2 December 2001, pp. 665-668.

[19] C. E. Shannon, "A Mathematical Theory of Communication", Bell System Technical Journal, Vol. 27, No. 3, 1948, pp. 379-423.

[20] T. M. Cover, J. A. Thomas, "Elements of Information Theory", 2nd Edition, Wiley, 2006.

[21] D. Freedman, R. Pisani, R. Purves, "Statistics", 4th Edition, W. W. Norton & Company, 2007.

[22] R. E. Walpole, R. H. Myers, K. Ye, "Probability and Statistics for Engineers and Scientists", 9th Edition, Pearson, 2011.

[23] IBM Almaden Research Center, "Quest Synthetic Data Generation Code and Datasets", http://cvs.buu.ac.th/mining/Datasets/synthesis_data/ (accessed: 2025)

[24] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C. Wu, V. S. Tseng, "The SPMF Open-Source Data Mining Library Version 2", Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, Riva del Garda, Italy, 19-23 September 2016, pp. 36-49.

[25] D. Dua, C. Graff, "UCI Machine Learning Repository", https://archive.ics.uci.edu/ (accessed: 2025)

[26] D. Fan, J. Wang, S. Lv, "Optimization of Frequent Item Set Mining Parallelization Algorithm Based on Spark Platform", Discover Computing, Vol. 27, No. 1, 2024, p. 38.

[27] M. Shaikh, S. Akram, J. Khan, S. Khalid, Y. Lee, "DIAFM: An Improved and Novel Approach for Incremental Frequent Itemset Mining", Mathematics, Vol. 12, No. 24, 2024, p. 3930.

[28] X. Sun, A. Ngueilbaye, K. Luo, Y. Cai, D. Wu, J. Z. Huang, "A Scalable and Flexible Basket Analysis System for Big Transaction Data in Spark", Information Processing & Management, Vol. 61, No. 2, 2024, p. 103577.

[29] S. Raj, D. Ramesh, P. Gantela, "CrossFIM: A Spark-Based Hybrid Frequent Itemset Mining Algorithm for Large Datasets", Cluster Computing, Vol. 28, 2025, pp. 231-245.

[30] Y. Rochd, I. Hafidi, "Frequent Itemset Mining in Big Data with Efficient Distributed Single Scan Algorithm Based on Spark", International Journal of Intelligent Engineering and Systems, Vol. 18, No. 2, 2025, p. 101908.

[31] A. Singla, P. Gandhi, "An Algorithm to Optimize Frequent Pattern Mining in Parallel and Distributed Environment", Engineering, Technology & Applied Science Research, Vol. 15, No. 3, 2025, pp. 22252-22256.

[32] T. Tassa, "Secure Mining of Association Rules in Horizontally Distributed Databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 4, 2014, pp. 970-983.

[33] R. Oliveira, O. R. Zaïane, "A Framework for Efficient Distributed Mining of Association Rules", Proceedings of the 2nd SIAM International Conference on Data Mining, Arlington, VA, USA, 11-13 April 2002, pp. 1-11.

[34] V. S. Tseng, C. H. Lin, Y. Y. Lin, C. W. Chen, "A Scalable Approach to Mining Frequent Itemsets over Big Data", Expert Systems with Applications, Vol. 42, No. 21, 2015, pp. 7876-7888.

Assessment of Battery Degradation Using Rainflow Cycle-Counting Algorithm: A Recent Advancement

Review Paper

Mohamad Faizal Yusman Mohd Hanappi
Department of Electrical,
Electronic and Systems Engineering,
Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi,
Selangor, Malaysia
p117837@siswa.ukm.edu.my

Ahmad Asrul Ibrahim *
Department of Electrical,
Electronic and Systems Engineering,
Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi,
Selangor, Malaysia

Advanced Power Engineering,
Centre for Automotive Research,
Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi,
Selangor, Malaysia
ahmadasrul@ukm.edu.my

*Corresponding author

Nor Azwan Mohamed Kamari
Department of Electrical,
Electronic and Systems Engineering,
Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi,
Selangor, Malaysia

Advanced Power Engineering,
Centre for Automotive Research,
Faculty of Engineering and Built Environment,
Universiti Kebangsaan Malaysia, 43600 UKM Bangi,
Selangor, Malaysia
azwank@ukm.edu.my

Mohd Hairi Mohd Zaman
Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
hairizaman@ukm.edu.my

Abstract *– Battery based energy storage systems are increasingly popular in power systems as renewable energy continues to grow while ensuring the reliability of power supply. However, battery degradation is a significant issue that can impact power system operations and optimal scheduling strategies. Therefore, estimating the remaining life cycle or assessing the health of batteries due to the degradation process has become a new challenge and research focus in various engineering fields. This topic is relevant in the context of electric vehicles (EVs), where battery degradation caused by continuous and non-continuous operations (i.e., charging and discharging cycles). Degradation can limit the performance of batteries and occur throughout their lifespan whether they are in use or not. The degradation process is complex and influenced by usage and external conditions that are normally measured by state of health (SOH). Therefore, predicting the SOH of batteries is crucial in ensuring the safety, stability, and long-term viability of energy storage and EVs systems. This prediction requires a battery mechanism model that can be established from a complex electrochemical process. Alternatively, a rainflow cycle-counting algorithm (RCCA) has become popular among researchers for battery degradation estimation because of its simplicity. This paper presents a comprehensive review of the battery degradation estimation using RCCA to count the equivalent cycles of charging and discharging profiles.*

Keywords: *Battery energy storage, Electric vehicles, Rainflow cycle-counting algorithm, State of health*

Received: March 11, 2025; Received in revised form: September 4, 2025; Accepted: September 10, 2025

1. INTRODUCTION

Renewable energy sources (RES) are one of the key solutions for global environmental pollution problems. However, renewable resources are intermittent, and their output heavily depends on weather conditions and local factors, thereby leading to new challenges especially in maintaining a good quality and reliable power supply [1]. In the last few years, the fluctuations in electricity generation from RES become a prominent issue among researchers and their main interest to solve. One effective solution is battery energy storage (BES) [2–5]. Cur-

rently, research on BES in power systems mainly covers the characteristics of system control [4], configuration modes [3], and mitigation actions [6, 7]. The operating cost due to degradation is a critical factor for the battery applications in power system. Therefore, extending the battery life cycle can significantly reduce maintenance and replacement costs [8]. The idea of using electric vehicles (EVs) as BES in the power system under the concept of vehicle-to-grid (V2G) was recognized as early as the last decade. The practicality of using EVs to provide ancillary services for power system including frequency regulation, base load fulfillment, peak shaving, and spinning reserve has been examined and tested, thereby leading to evaluations of the economic benefits when using different technologies such as battery, fuel cell, or hybrid plug-in vehicles [9].

EVs have attracted global attention due to their energy efficiency and environmentally friendly features. With the rapid advancement of technology, the use of batteries in the automotive sector has also become increasingly popular. As a result, the performance of rechargeable batteries presents a key concern for users. Furthermore, the cost of batteries contributes up to 30% in manufacturing of an EV, thereby limiting the development of EVs [10, 11]. Lithium-ion (Li-ion) batteries are preferrable for EVs as compared to other types due to their superiority in terms of performance, size, weight, and impact on environment [12]. However, the main concern of EV applications using Li-ion batteries are their safety and reliability. Poor road conditions, temperature changes, and load fluctuations can degrade the batteries performance when they are used outdoors. Apart from that, the performance degradation can be caused by insulation failure, current leakage and short circuits, and if not addressed appropriately and timely, can result in serious incidents, such as spontaneous combustions and explosions [12-15]. For that reason, monitoring the performance and estimating the degradation of batteries based on their state of health (SOH) are of great concern for EV users. Measuring the health or feature of batteries at their current state are required before estimating their SOH.

Three SOH estimation approaches are commonly used for Li-ion batteries, namely, the battery impedance method, ampere hour counting method, and cyclic method [16-18]. The cyclic method, which will be the main focus of this review, is based on a simple principle that does not require various measuring parameters [19]. However, this method requires high accuracy in monitoring the number of cycles. While previous studies on SOH primarily focused on impedance and capacity measurements, only few explored the life cycle of batteries. Furthermore, in-depth discussions on effective cycle criteria are limited due to the strong nonlinear characteristics of batteries, with most studies relying on impedance and capacity criteria [20]. Therefore, the limitations of the cyclic method in accurately estimating the number of cycles remain unsolved. This paper reviews the applications of the Rainflow cycle counting algorithm (RCCA), which is one of the cyclic methods used to establish degradation models, in assessing SOH. The main contributions of this paper as the following:

- A comprehensive review on RCCA to estimate an equivalent cycle for battery degradation assessment in electric vehicle and power system applications.
- The detailed calculations of the conventional and improved RCCA for better understanding on the equivalent cycle's computation for battery degradation assessment.

The remaining of this paper is organized as follows. Section 2 explains the concept of battery life cycle assessment. Section 3 discusses the applications of RCCA in power systems and electric vehicles. Section 4 outlines the improvements to RCCA. Section 5 draws a conclusion for this paper and provides a direction for future works.

2. BATTERY DEGRADATION ASSESSMENT

SOH is defined as a ratio between the maximum discharging capacity of the batteries and their nominal capacity. Given that the maximum discharge capacity is a characteristic of battery aging, SOH is used as an indicator of the degree of aging. A new battery without any degradation has an SOH of 100%. Fig. 1 depicts a general SOH curve over time to visualize the degradation based on the expression in (1).

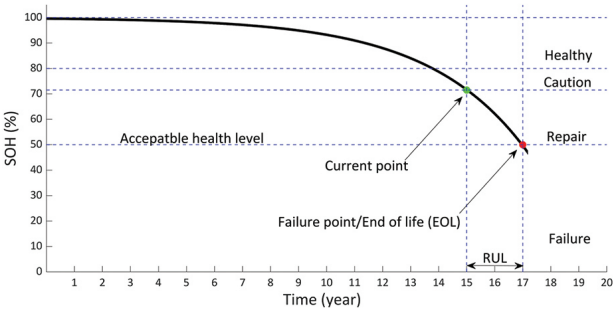


Fig. 1. A general SOH estimation curve [21]

$$SOH = \frac{Q_i}{Q_{max}} \times 100\% \quad (1)$$

where, Q_i refers as the battery capacity in Ampere-hour at i-th cycle, while Q_{max} is a new or fresh battery capacity in Ampere-hour. The degradation curve varies and depends on the battery type and characteristic. Furthermore, the battery can be replaced, or cautious action can be taken when the SOH reaches approximately 70%. Meanwhile, remaining useful life (RUL) is measured from the caution point (~70% SOH in the figure) until the SOH reaches the failure point or end of life [22]. A preventative or predictive maintenance procedure is useful to reduce battery failure rates and maintenance costs.

Battery life cycle assessment is an important step to determine cost over the reliability of battery-powered

devices. Numerous life cycle models have been discussed in the literature, but they often require a compromise between precision and generality. Some models use a generalized equation which is derived from experimental data to evaluate the relationship between lifetime and relevant parameters such as depth of discharge (DOD). The relationship of battery degradation to its cycle life was also studied, and one of the significant factors that determine degradation is the number of cycles. Degradation is caused by irreversible physical and chemical changes within the battery, most commonly occurring during charging or discharging. The most visible sign of deterioration is a decrease in battery capacity after repeated charge and discharge cycles. However, the models tend to have low accuracy, and the relatively accurate models that account for the effect of aging in an equivalent circuit are usually specific to the battery being tested and are not generally applicable. A new estimation model of battery cycles taking into account capacity loss is proposed in [23]. This model can effectively explain the cycling behavior of batteries at different chemical compositions and able to make accurate battery life cycle assessment.

A reliable life cycle model is essential to accurately estimate the battery's SOH during operation and ensure demand is met. Motapon *et al.* [24] has developed a hardware testbed to evaluate the cycle-based aging process of a Li-ion battery and its impact on the battery's internal resistance and capacity. This model is based on fatigue theory and equivalent cycle counting that requires only limited data from battery data sheets and short-term cycle experiments to identify the relevant parameters. However, as batteries near end of life (EOL), they are prone to instability during charging or discharging and other problems such as overheating and excessive current. The temperature estimation method contains elements that could lead to errors in the battery SOH analysis, including errors in temperature measurement and in the predicted dynamic characteristics of temperature changes due to the state boundary. Kim *et al.* [25] introduced an effective approach to predict the life cycle of Li-ion batteries based on the entropy law and obtained promising results. Although temperature and time functions play an important role in the estimation, voltage and current are more responsive and effective in real-time battery state acquisition and processing. Adermann et la. [26] proposed a commuter cycling monitoring model to estimate the parameters of EV batteries. This model benefits from simple algorithms that work effectively with only a few measurements, enable real-time application on vehicle hardware, or outsource the assessment to a back end for advanced data gathering and processing. However, this model also requires a close relationship between the battery state of charge (SOC) and open circuit voltage (OCV) and assumes extensive knowledge of their relationship. This obstacle makes the application of this approach difficult because SOC is also influenced by other factors such as temperature.

An optimal scheduling strategy of renewable resources and BES in microgrid (MG) was used in [27] to minimize energy costs based on forecasted data of renewable energy generation, electricity prices and electricity demand. The most useful BES measurement in this case is to monitor the active power transferred back into the power grid. The costs due to battery life cycle were also taken into account and a recursive cost model was developed. Battery life cycle estimation is also crucial for designing solar home systems (SHS) and it requires experimental data to model the electrochemical processes of a battery at the cell level. Narayan *et al.* [28] developed a practical approach to estimate battery life cycle without having to perform experimental works or model the electrochemical processes in the battery. This method is based on battery data sheet provided by manufacturers. Therefore, it does not rely on technology-specific electrochemical processes where it can be used in other battery applications subjected to similar characteristics without affecting its accuracy. A summary of the discussed battery life cycle assessment approaches is presented in Table 1.

Table 1. Review of battery degradation assessment approaches

Research work	Highlight/Advantage	Limitation
[23]	Applicable for various chemistries	Temperature and current rates are fixed
[24]	Degradation model parameters are simplified	Limited to certain types of batteries
[25]	Model based on entropy law	Effect of temperature on degradation is not significant
[26]	Fast data collection and processing	SOC is derived from open circuit voltage only
[27]	Compatible with dynamic programming	An additional analytical approach is required
[28]	A dynamic capacity fading	Low C-rates are neglected

3. RAINFLOW CYCLE COUNTING ALGORITHM

RCCA was introduced by Matsuishi and Endo in 1950 to calculate a fatigue life and represent it in a load-time curve that enables the measurement of the actual stress history over several cycles of damage accumulation [29, 30]. RCCA uses SOC profiles to estimate battery life cycle. A SOC versus time graph is plotted and rotated 90 degrees clockwise where time axis is pointed vertically downward as depicted in Fig. 2.

In the figure, valleys are labeled with a, c, e, g, i, and k, while peaks are labeled with b, d, f, h, j, and l. The dropping rainflows at these peaks and valleys are denoted by DP and DV, respectively while, the rainflow between the peak and valley (i.e., on the rooftop) is denoted by RF. The following rules are observed in RCCA:

- A rainflow between a peak and valley (RF) starts at each peak or valley and stops at the opposite end point if there is no obstacle in between.

- A dropping rainflow is created after the RF reaches the end point. In this case, DP or DV is created after the RF reaches a peak or valley, respectively. New dropping rainflows are created at all valleys and peaks except those from the dropping rainflows at e and i.
- The dropping rainflow stops when the next dropping point is greater (i.e., smaller for valleys or higher for peaks) than or equal to the previous point just before the dropping is created. In this case, DP stops when the next valley is smaller than or equal to the previous valley or DV is stopped when the next peak is higher than or equal to the previous peak. For example, DP from peak b stops when the valley at c is smaller than the previous valley at a. Meanwhile, the DP from peak f continues when the valley at g is higher than that at e and keeps continuing when valley at i remains higher than that at e.
- The rainflow stops upon meeting another rainflow. In this case, the RF stops upon meeting the dropping rainflow (either DP or DV; for instance, at c', g', and j') and is replaced with the respective rainflow (i.e., DP or DV, respectively). This case also applies when the dropping rainflow meets an earlier dropping rainflow (for example, at f'), and the current dropping rainflow is replaced with the earlier dropping rainflow.

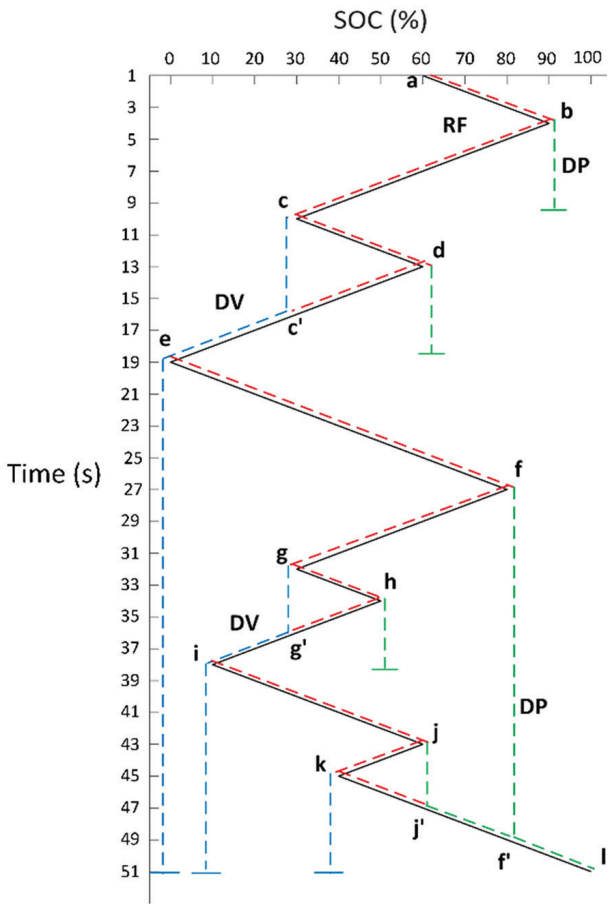


Fig. 2. A rotated curve of SOC over time

A half cycle is counted whenever the RF stops, and a full cycle is counted when the RF meets the DP or DV while considering a starting point from where DP or DV is created. Fig. 2 shows 3 full cycles for c-d-c', g-h-g', and j-k-j' and 5 half cycles for a-b, b-c, e-f, f-g, and i-j. Therefore, a total of 5.5 cycles is counted, and their amplitudes are recorded for further analysis. Applications of RCCA for battery life cycle assessment can be divided into power systems and EVs as will be discussed in the following subsections.

3.1. APPLICATIONS IN POWER SYSTEMS

A large-scale BES is normally used in power systems to improve their operation. Muenzel *et al.* [31] developed a battery life cycle prediction technique that focuses on the operational optimization of battery management. This technique considers multiple changing cycling parameters of Li-ion battery cells. Five operating factors are considered in four separate models, including charge and discharge currents, maximum and minimum operating cycle limits, and temperature. The models were then calibrated using experimental battery data. RCCA and discretization technique were used to incorporate dynamic factors into the battery cycle profiles and to solve the optimal battery operation problem. Xu *et al.* [32] introduced a semi-empirical degradation model to make assessment on the battery cell degradation from its charging and discharging profiles. This model can be applied for various types of Li-ion batteries by tuning the model based on manufacturer's data. The incorporation of RCCA allows the model to determine stress cycles from irregular charging and discharging operations. Shi *et al.* [33] established a convex RCCA degradation cost with respect to BES charging and discharging operations. The degradation model can be easily incorporated into a sub-gradient of the optimization algorithm due to its convexity. Shi *et al.* [34] later proposed an optimal control of BES to maximize profit under a "pay-for-performance" scheme where a payment is made when the BES operation complies with the utility company instructions. The degradation cost was also considered by using the convex RCCA as discussed in the previous work.

Assessing the economics of using batteries to reduce peak demand and price arbitrage is becoming attention among researchers and energy suppliers. Schneider *et al.* [35] developed an approach to minimize investment and operational costs by determining appropriate battery technology and size, and scheduling the battery operation, respectively. RCCA has been integrated into a multitasking optimization platform for battery selection and shipping. The results of a simulation study conducted by a Swiss power provider showed that battery integration can make economic sense if its capacity and drive units are carefully selected, highlighting the importance of battery size selection. Rosewater *et al.* [36] presented an advanced optimal control method to

maximize the benefits of battery integration. The SOC, temperature and SOH of Li-ion battery cells are modeled in a predictive controller that allows battery operation scheduling, air conditioning and forced air convection, optimizing energy consumption and reducing electricity bills. RCCA was also used to develop the SOH model that produces a linear relationship between battery usage and degradation. Singh *et al.* [37] proposed Mixed Number Linear Programming (MILP) to optimize the operation of home appliances and manage energy from distributed energy resources (DERs) and the power grid based on price-based pricing and usage hours. An energy management system and a load planning system were developed, integrated into a house. Data were analyzed using RCCA to assess the decline in performance of residential BESs through EOL.

The financial benefits of BES are usually estimated based on the profits gained from system operations by utilizing batteries, but this approach ignores the fact that battery operations reduce the battery lifetime itself. Foggo *et al.* [38] developed a framework for BES valuation that co-optimizes with a realistic degradation model to maximize profits and mitigate battery degradation at the same time. RCCA was used to calculate equivalent cycles from SOC profiles that gives the battery degradation. Lee *et al.* [39] proposed a new battery degradation cost formula for optimal BES operation planning. The RCCA-based mining cost was formulated as piecewise linearity using an auxiliary SOC. Therefore, the optimal scheduling of BES together with the battery life cycle characteristic can be modelled as a MILP problem and solved using a gradient-based solver. As a result, an optimal scheduling BES operation can be determined quickly. Soleimani *et al.* [40] presented an active distribution network (ADN) that uses an energy storage system (ESS) within their constraints to optimize battery lifespan and minimize the operating cost. They found that BES is operated at a lower rate if the battery lifespan is taken into account, thus underutilizing the battery capacity. In their later work, Soleimani *et al.* [41] proposed a method for BES scheduling in ADN to minimize operating costs and reduce the impacts on BES lifespan. This method uses a linearized and convex AC-OPF model for a quick and accurate calculation. A two-stage stochastic optimization approach and K-means clustering were also used to address the uncertainties in different case scenarios where the battery degradation was determined using RCCA.

The penetration of DERs has created challenges in distribution network to maintain its operation within the statutory limits. Tang *et al.* [42] proposed a Lagrangian-relaxation-based algorithm that solves an optimal BES scheduling in distribution networks with DERs by incorporating an RCCA-based degradation model and using Copula theory to capture the uncertainties of DERs. This algorithm allows the incorporation of more scenarios into the BES scheduling framework and effectively captures the uncertainties of DERs. Chawla *et al.* [43] examined the major applications of energy storage in utilities as well as the requirements and challenges faced by BESs. RCCA was used to estimate the battery degradation under dynamic duty cycles, assuming that duty cycles are known in advance and that battery degradation in microcycles is independent of macrocycles. This work illustrates the trade-off between the initial investment cost of BESs (i.e, battery sizing) and the battery life cycle degradation cost. Abdulla *et al.* [44] introduced a stochastic dynamic programming approach that optimizes ESS performance over a shorter time horizon by leveraging available forecasts and a multifactor battery degradation model that takes operational influences into account. This approach aims to maximize the battery life cycle based on information from the forecasted data and operational impacts on battery degradation. This degradation model uses a dynamic RCCA that uses the time-history of discharge profiles to determine the equivalent degradation cycles.

Photovoltaic (PV) energy production fluctuates due to high intermittent in the solar radiation intensity caused by moving clouds. An ESS equipped with a ramp-rate (RR) control can be used to mitigate the fluctuations of PV output. Martins *et al.* [45] conducted a comprehensive analysis of PV power balancing techniques using ESS through RR control scheme and ensure SOC at the end of the day is remain as the start. ESS capacity requirements were quantified using RCCA from operation profile and DOD analysis. A grid-connected PV system aims to generate power according to the hourly production bids in the electricity market to avoid penalties. Beltran *et al.* [46] analyzed the aging of six different battery chemistries, including Li-ion, Sodium-sulfur, Nickel-cadmium, Nickel-metal hydride, Lead-acid, and Lead-gel, in a large-scale grid-connected PV system that participating in the electricity market. A systematic annual analysis was performed using RCCA to determine the number of cycles experienced by BES. In this case, BES was used to ensure that the energy input from PV meets the market demand. Alam *et al.* [47] elucidated the influence of PV variability on the ESS life cycle using RCCA. A realistic concept of life cycle degradation was derived from data from a real PV system in Australia. Hossain *et al.* [48] implemented a preventative energy management scheme that taking into account the battery degradation costs to accurately represent the actual cost of ownership. The management scheme considers the operation cost of battery from charging/discharging profiles and then, uses particle swarm optimization and RCCA to minimize the cost.

As RES become more widespread, the need for fast and reliable support services increases. Ochoa-Eguilegor *et al.* [49] analyzed ESS's participation in dynamic storage and continuous intra-day auctions in the UK. A battery SOC management strategy was developed, and the battery life cycle was estimated using an aging model based on RCCA and Wöhler curves. Furthermore,

a techno-economic analysis was carried out to demonstrate the technical feasibility and reliable operation of the BES. Karmiris *et al.* [50] evaluated different control methods for BES in renewable power smoothing applications. The effectiveness of each control algorithm in terms of renewable smoothing and battery stress was analyzed, and the battery stress and life cycle were estimated using RCCA. A good renewable smoothing strategy can negatively impact the battery life cycle. Bouakkaz *et al.* [51] proposed a strategy for maximizing battery life cycle by managing the battery operations. This strategy minimizes the number of battery cycles per day by scheduling adjustable loads and controlling the charging and discharging processes. The optimization problem was solved using particle swarm optimization, and RCCA was used to calculate the number of battery cycles. Dragicevic *et al.* [52] proposed a technique to minimize the energy consumption of an autonomous remote installation based on robust mixed-number linear programming. This model identifies the optimal combination of renewable energy and ESS, considering the service life of the telecommunications system and the attractiveness of different battery technologies. This technique shows flexibility in solution accuracy and computational load, and RCCA was applied to account for the DOD-related cycles.

Lee *et al.* [53] presented an optimal scheduling framework for BES in MG to address the uncertainties in RES and load demand. This framework minimizes BES service life degradation and ensures economic viability of MG operations. Monte Carlo simulation and K-means clustering algorithm were used to deal with the uncertainties, while RCCA was used to process the BES charging/discharging profiles. In isolated MGs, the integration of RES, diesel generators and storage batteries are necessary to minimize fuel consumption and ensure continuous power supply. Boqtob *et al.* [54] investigated the optimal power distribution for MG engines and used RCCA to count charge/discharge cycles and quantify battery degradation. Li-ion batteries are widely used for real-time power balancing to ensure the economic operations in islanded MGs. With respect to battery degradation, Lyu *et al.* [55] proposed a novel degradation model for Li-ion batteries in islanded MGs that considers real-time management using RCCA and an online auction system. This model was formulated as a mixed integer non-linear programming (MINLP) and used weighted model predictive control to address uncertainties in the look ahead window.

The future of smart grids highly depends on large battery storage. As the use of batteries in energy markets continues to increase, the need for an optimal bidding strategy becomes increasingly important. Batteries can increase profitability through a rapid regulation service based on their performance. However, frequent charge/discharge cycles can shorten battery life, especially with quick setup services. He *et al.* [56] developed an auction model that takes battery life cycle into account for profit maximization in the energy market bidding. This model determines optimal bids in energy, reserve trading and day-ahead regulatory markets and uses an online distributed calculation method to decrease its complexity. The model offers battery storage investors a valuable tool to make decisions about tenders and operating programs and to assess economic feasibility. Correa-Florez *et al.* [57] proposed a stochastic approach for home energy management systems (HEMS) that considers BESs, PV resources and electric water heaters in daily operation framework. A swarm optimizer minimizes operating costs by considering the purchase of energy from the wholesale market and the corresponding cost of battery aging. This approach takes into account uncertainties in PV production and charging and is a valuable tool for optimizing HEMS operations. The cost of cyclic battery aging was considered using a memory disaggregation algorithm based on Lagrangian relaxation and RCCA. This approach can handle complex switching behavior and reduces the search space in optimization problem. Therefore, the decomposition strategy is supplemented by a competitive swarm optimizer.

Rapid charge/discharge operation in off-grid wind energy systems and high discharge currents during motor start-up and other high-load scenarios can reduce the battery life cycle. Li *et al.* [58] attempted to solve this issue by integrating superconducting magnetic energy storage into conventional batteries to minimize short-term power cycling and high discharge currents. The wind power was incorporated with ESS, load fluctuations and wind turbulence to demonstrate system performance. A battery life model was also used to estimate the improvement in battery life due to the reduction of charge/discharge cycles and discharge rate, while RCCA was used to isolate irregular charge cycles and discharges experienced by the battery within the simulation period. Pan *et al.* [59] integrated cyber-physical systems (CPS) into the control framework of hybrid energy storage system (HESS) and used multi-objective optimization to solve the problem. The performance of HESS can be significantly improved by incorporating physical models and real-time data through CPS. The multi-objective optimization control scheme was developed for the HESS battery supercapacitor that considers component characteristics, reduces power consumption and maintains SOC within the required limits. RCCA was used to predict battery life and quantify the benefits of using HESS as part of the control plan.

Open cycle gas turbines (OCGTs) are often used to provide fast-frequency regulation in maintaining the system frequency within the required limits. However, these OCGTs are expensive. As an alternative, frequency regulation can be provided using ESS due to its quick response capability. Lian *et al.* [60] proposed a suitable size of OCGTs and ESSs to provide frequency regulation in response to load fluctuations. RCCA was

used to determine the battery life cycle, and to accurately determine the cost savings from ESS specifically for frequency regulation, hence highlighting the advantages of ESS over OCGTs. Loew *et al.* [61] implemented a cycle identification system using RCCA in model predictive controller for Li-ion batteries to accurately estimates the revenue of ESS by considering the cost of aging. Anand *et al.* [62] improved the large-scale integration of wind energy into the power grid. An economical nonlinear model predictive controller (ENMPC) was developed to operate a wind turbine and a battery as a hybrid system to supply energy to the grid. ENMPC calculates the revenue from electricity generation considering the costs associated with mechanical fatigue damage to the wind turbine tower and the cyclical loss of Li-ion battery capacity. An online parametric RCCA was implemented to determine the cyclical loss.

Traditional energy generation facilities have exhibited a marked dependence on hydrocarbon resources to meet the increasing energy requirements prompted by accelerated demographic expansion and an array of technological advancements. Obaro *et al.* [63] modelled an optimal energy framework and power management strategy for an off-grid distributed energy system (DES). The management strategy is co-optimized with various energy generation modalities as a fundamental objective to guarantee reliable and cost-effective power delivery to electrical loads, whilst conforming to a defined set of operational system requirements. Furthermore, the MINLP optimization methodology is applied to improve the generation efficiency of power systems that are interconnected with diverse energy sources and variable electrical demands. Considering the recurrent cycling behavior of batteries within the DES, the RCCA is implemented to calculate the cumulative number of cycles.

The cost function of Li-ion battery within the electricity market necessitates an optimal equilibrium between the maximization of revenue derived from energy arbitrage and the minimization of capacity degradation resulting from operational usage. The optimal equilibrium can be attained by integrating the stresses associated with DOD and thermal conditions of the battery into the optimal economic dispatch framework. A series of physics-based sufficient conditions have been formulated to effectively manage the non-analytical nature of RCCA, while simultaneously accounting for temperature variations at the cell level. The suggested stress-conscious optimal battery dispatch (SC-OBD) paradigm is executed within the framework of a battery operating in both day-ahead and real-time balancing market environments. Furthermore, Singh *et al.* [64] introduced a predictive control-based framework model to address the unpredictability of real-time electricity pricing and the need to guarantee adherence to agreements made in the day-ahead market. Table 2 summarizes the applications of RCCA for battery degradation assessment in power systems.

Table 2. RCCA applications in power systems

Research work	Highlight/Advantage	Limitation
[31]	Multiple changing cycling parameters are considered	A fixed operating temperature
[32]	Adaptable degradation model to various Li-ion batteries	Extreme conditions (i.e, low SOC, over-voltage, etc.) are not considered
[33]	A convex degradation model	Underestimate the actual degradation effects
[34]	An online model for market bidding	Uncertainties of loads are not considered
[35]	An optimal sizing of battery for investment	A calendar aging is neglected
[36]	A linear degradation relationship with battery usage	A specific usage behavior (air-conditioning)
[37]	A smart residential energy management system	Open loop battery capacity assessment
[38]	An improved battery lifetime framework	A pre-determined set of battery actions
[39]	A piecewise linear degradation cost	High computational cost
[40]	Battery utilization in power grid ancillary services	A specific type of batteries
[41]	A linear and convex AC-OPF model	High C-rates are neglected
[42]	A correlation between degradation cost and DER uncertainties	The potential voltage instability of DERs is underestimated
[43]	Battery degradation accounted in micro cycle	Internal battery resistance is neglected
[44]	A multi-factor battery degradation model	A time value-of-money is not accounted
[45]	Constraints on RR and SOC endpoint are included	Battery capacitance is neglected
[46]	Various battery chemistries for a large-scale PV	Operating temperature is fixed
[47]	A realistic life cycle degradation for PV plant	Low DOD is not considered
[48]	An optimal energy management for PV and ESS system	High computational time
[49]	Wöhler curves are integrated for aging evaluation	Variation of DODs is not considered
[50]	Various battery stress conditions including low DOD	Variation of C-rates is neglected
[51]	Optimal scheduling of shiftable loads and battery operation	Intermittent of resources and loads is not considered
[52]	Lifespan of the telecommunications facility is considered	Time consuming due to integer variables
[53]	A Monte Carlo simulation is used to consider uncertainties	Applicability to real system is not tested
[54]	Battery degradation in optimal energy dispatch framework	Tested on a specific type of batteries
[55]	Energy market bidding during an islanded operation	Irregular life cycle profiles
[56]	A fast market bidding decision for battery owners	Appropriate size of battery is not identified
[57]	Lagrangian relaxation is used for model simplification	Limited for planning perspective
[58]	Integrated with superconducting magnetic energy storage	Effect on different temperature is ignored
[59]	Actual field data is considered using CPS platform	Multiple units require additional scheduling framework

[60]	A cost saving of ESS for frequency regulation	Huge deviation between actual and projection loads
[61]	A moving horizon of cyclic aging based on MPC	Not tested on real-life battery storage
[62]	Non-linear characteristics of degradation are considered	Micro cyclic damage is not considered
[63]	A modular multi-energy sources to improved power reliability	Micro cycling operation is neglected
[64]	A stress-aware optimal battery system dispatcher	Low DOD C-rates are neglected

3.2. APPLICATIONS IN EVS

EVs primarily rely on batteries for operation, necessitating an effective management system to enhance their performance. Muenzel *et al.* [31] developed a battery life cycle prediction technique that focuses on the operational optimization of battery management. The life cycle discussions presented in the previous subsection are mainly based on the planning perspective, which uses a large time interval spanning 15-60 minutes. However, the battery charging and discharging profiles for EVs have a much shorter time interval that depends on the driver's action to accelerate (discharging) or decelerate the vehicle through regenerative braking (charging). In 2010, the Racing Green Endurance project designed and built the world's largest EV with a range of over 514 km [65]. Operational data of battery usage from the Racing Green Endurance project was used to develop a new battery degradation model. This model uses the RCCA method to produce highly reliable and accurate predictions of capacity and power losses in vehicle traction batteries. Li *et al.* [66] introduced a CPS-based electric vehicle platform to collect and store battery consumption data in the cloud, which can be used for battery degradation assessment. Support vector regression algorithm and RCCA were used to develop a battery degradation model and study the dynamic characteristics of the batteries. In their subsequent work, Li *et al.* [67] used RCCA with a deep learning algorithm to estimate the aging of EV's batteries. The RCCA-based approach effectively extracts the aging history of the battery and provides an aging index to evaluate the degradation.

EVs are becoming popular for public transportation because of target to reduce greenhouse gas emissions. Bai *et al.* [68] proposed a hierarchical optimization of energy management strategies using HESS to reduce the impact of battery aging in plug-in hybrid electric buses. This strategy includes a power limit management module that controls the flow rate of the supercapacitor and battery by redistributing power between them to drive the motor. A simple but effective battery life cycle model based on RCCA was used to quantify the rate of degradation in a battery performance control strategy. In another work [69], a low profit margin was identified as the main challenge in managing a fleet of EVs. A control strategy was proposed for managing a fleet of EVs in terms of charging and discharging for grid ancillary services. This strategy minimizes the operational cost by simultaneously determining the wear of batteries in EVs and assigning suitable routes for the ancillary services. The wear of battery was calculated using RCCA and the integral of the wear density function. Sandelic *et al.* [70] proposed an incremental degradation cost using RCCA to allow for a real-time evaluation of the true cost of battery operation while considering the degradation effects in a specific time interval.

An aggregator can achieve frequency regulation by controlling its generation and demand to cater the fluctuations in the electricity market due to increasing contributions from RESs. Vatandoust *et al.* [71] studied the participation of an aggregator to manage a fleet of EVs and ESS operations in a day-ahead energy market regulation framework. The fleet of EVs and ESSs provide unidirectional (charging) and bidirectional (charging and discharging) regulations, respectively. Risk-free mixed-integer stochastic linear programming was then applied to plan aggregator participation, and RCCA-based linear degradation was formulated to account for the expected degradation costs incurred by EVs participation. The degradation cost is a key concern for EV owners that discourages their participation in vehicle-to-grid (V2G) services for regulation purposes. Li *et al.* [72] proposed a novel anti-aging V2G active battery planning approach, which quantified battery degradation using RCCA during V2G services. The V2G scheduling problem is modelled as a multi-stage optimization problem that aims to minimize battery degradation and load fluctuations in the power grid.

In contemporary discourse, the implementation of sophisticated charging paradigms and management frameworks that encompass vehicle-to-everything (V2X) functionalities is essential to alleviate the increasing ubiquity of battery electric buses (BEBs) in urban settings. Nevertheless, the integration of these advanced functionalities into charging systems may have repercussions on the longevity of the charging infrastructure. This phenomenon remains unexplored, despite its significance for operators of BEBs. Verbrugge *et al.* [73] developed a thorough evaluation of reliability to investigate the consequences of intelligent and bidirectional (V2X) charging on the longevity of silicon carbide-based high-power off-board charging infrastructure employed for battery electric buses (BEBs) within a depot environment designated for overnight charging. The thermal stress is converted into a quantifiable metric of failure cycles and cumulative damage through the utilization of RCCA, a life cycle prediction model for damage accumulation. Ultimately, a Monte Carlo simulation along with a Weibull probability distribution fitting is utilized to determine the reliability of the system.

An essential improvement in the accessibility of rapid-charging infrastructure is crucial for the effective shift towards EVs. Nevertheless, the process of charging a battery pack at increased C-rate has detrimental effects on SOH, thereby expediting its deterioration. Pelosi *et al.* [74] proposed a strategic battery management, which

considers the diurnal operational patterns of a Li-ion battery utilized in EVs, anchored in a defined driving cycle that includes charging phases occurring when DOD attains 90%. Through the dynamic modeling of the EV's battery system, the progression of the state of charge is determined for a range of charging C-rates, with meticulous attention given to both discharging and charging profiles. RCCA was employed to examine the SOC profiles, thereby determining the DOD for each individual cycle, which subsequently informs the practical applications on the experimental testing apparatus. The above applications of RCCA for assessing battery degradation in EVs can be summarized in Table 3.

Table 3. RCCA applications in EVs

Research work	Highlight/Advantage	Limitation
[65]	Long-range empirical EV data	A specific battery type
[66]	An online SVR-based assessment	Inadequate data for verification purposes
[67]	A deep learning-based assessment	A fixed temperature and specific type of battery
[68]	An integrated with super-capacitor for electric buses	Maintenance cost is not considered
[69]	EVs scheduling for grid ancillary services	The wear function is not validated on actual batteries
[70]	An incremental degradation cost of EVs	A calendar aging is neglected
[71]	EVs participation in the electricity market regulation	Framework is mainly based on linear approximations
[72]	Minimize degradation during V2G services	Prediction errors of battery capacity fade are neglected
[73]	Silicon carbide (SiC)-based high-power off-board charging	Not tested on real-life battery storage
[74]	Capacity degradation tailored with high C-rates	Low C-rates are neglected

4. IMPROVEMENT AND FUTURE WORKS

Despite its wide usage in cycle counting, RCCA has a non-closed form that hinders its use in optimization. As a result, recent studies have explored ways to improve RCCA. For instance, Huang *et al.* [75] proposed an accurate life cycle prediction for Li-ion batteries. The charge and discharge profiles are usually faced with interference from noise, which is not addressed properly in the traditional ampere hour approach. The unscented Kalman filter algorithm addresses this issue and provides a highly accurate SOC for SOH prediction. Li-ion batteries also have a strong non-linearity characteristic that leads to many small cycles counted in traditional RCCA. An improved RCCA was then introduced by adding intermediate judgement to reduce its sensitivity to data peaks. A linear damage criterion was then used together with the improved RCCA to accurately predict the remaining life of a battery without the need to measure process parameters. In a later study, Huang *et al.* [76] improved RCCA by combining this approach with the autoregressive integrated moving average (ARIMA) model to predict the SOH of a Li-ion battery. Experiments were conducted under dynamic stress tests and cycle conditions to validate the performance of the SOH prediction model using a confidence interval as the acceptable error range. The combined RCCA and ARIMA show promising results in predicting the SOH of batteries.

BES duty cycle counting in frequency control is hard due to irregular charging and discharging caused by fluctuations in grid frequency. The traditional RCCA is based on extreme points (peaks and valleys) and only starts counting at the end of data, hence it is inapplicable to determine the RUL in between load points, especially in real-time applications. Gundogdu *et al.* [77] modified the RCCA to develop a rapid/fast battery cycle counting method that estimates an equivalent number of completed cycles that can be used to calculate RUL in microcycles. Unlike traditional RCCA, the proposed method can calculate half a cycle when the SOC of each battery charge and discharge independently reaches the maximum value of 100% and one full equivalent cycle can be achieved as single battery charge and discharge cycles are recorded. Furthermore, the number of complete equivalent cycles can also be estimated during the process continuously rather than waiting for the data collection to end. This method was applied to 1 MWh BES at 2 MW maximum operating power to mitigate the frequency fluctuation problem. The lack of a comprehensive mathematical formulation for the RCCA have represented the one of primary barriers to the widespread implementation of the cycle-based degradation framework. Diao *et al.* [78] proposed a meticulous analytical formulation of the sub-gradients pertaining to the cycle-based aging cost function to facilitate the effective resolution of the optimal operational dilemma irrespective of a mathematical representation for the RCCA. The sub-gradient projection algorithm is introduced to determine the theoretical optimal operation in particular circumstances where the constraints governing battery operations may be alleviated.

An example application for the improved RCCA is also provided in this review for further understanding. Fig. 3 depicts a flowchart of the improved RCCA used for real-time applications.

The same notations presented in the previous section are used in the figure. A flag is used to indicate the condition wherein RF meets DV or DP where (1 if true, and 0 if false). The direction of rainflow is important in determining whether the next edge is a peak or valley. Therefore, $DirVP$ is used to indicate a flow from valley to peak ($DirVP = 1$) or, from peak to valley ($DirVP = 0$). In a casual event as mentioned in the first rule in the rainflow cycle counting algorithm section, an amplitude (Amp) of the half cycle is calculated as:

$$Amp = P - V \tag{2}$$

In the event where RF meets DP or DV (flag = 1), Amp is recorded using the following expression:

$$Amp = \begin{cases} DP - V, & \text{if } DirVP = 1 \\ P - DV, & \text{otherwise} \end{cases} \tag{3}$$

The equivalent number of cycles can then be updated using the recorded Amp at each time step in as follows:

$$Cycle(t) = Cycle(t-1) + 0.5\left(\frac{Amp}{100}\right) \quad (4)$$

Fig. 4 shows the total number of equivalent cycles in respect to SOC at each time step using the improved RCCA algorithm. The SOC curve in Fig. 4(a) is based on data as in Fig. 2. The total number of equivalent cycles at the end of the data can be observed at approximately 2.1 cycles as shown in Fig. 4(b). The equivalent cycle is much smaller than that obtained by the traditional RCCA at 5.5 cycles as discussed earlier.

The equivalent cycle plot in Fig. 4(b) resembles a staircase due to the cycle cannot be updated until the SOC reaches peaks or valleys. A plateau at the end of the equivalent cycle plot is caused by the failure to locate a peak or valley. The limitations of equivalent cycle counting should be addressed in the future works to ensure a smooth and accurate representation of the battery degradation process.

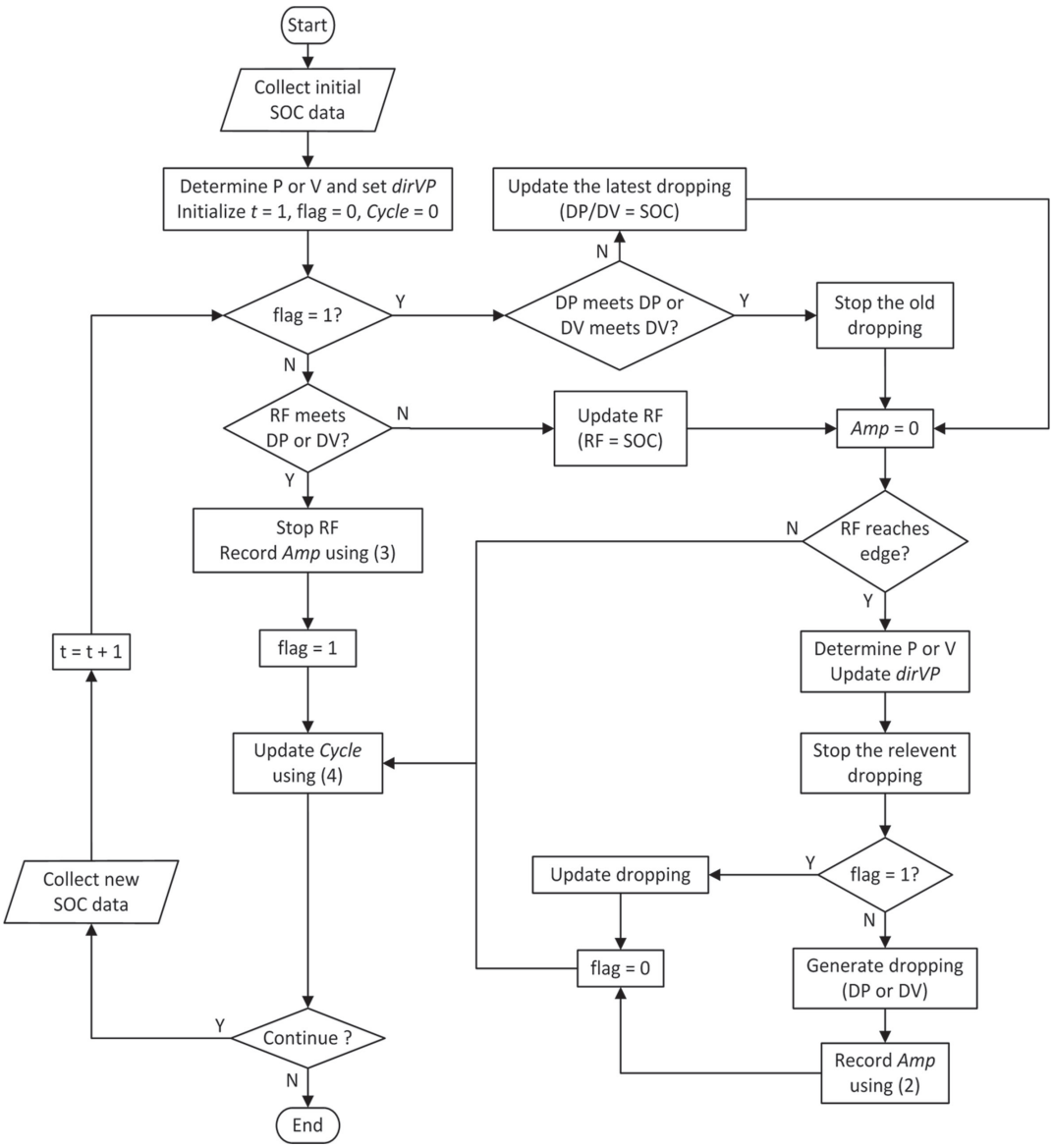
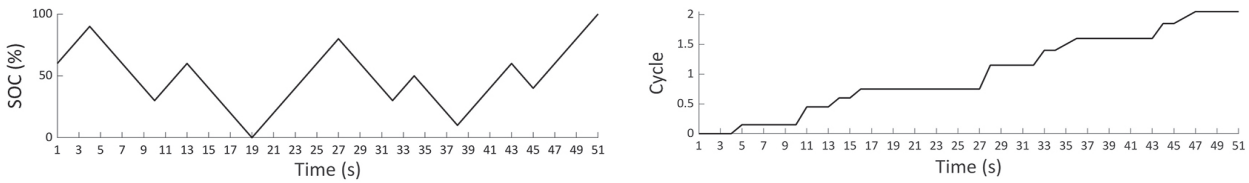


Fig. 3. Flowchart of the modified RCCA



(a) (b)

Fig. 4. Equivalent cycle counting using the improved RCCA algorithm

5. CONCLUSIONS

This paper discusses a comprehensive overview of the applications of RCCA in battery degradation assessment. The relevant literature on the topic is reviewed by highlighting the advantages and limitations of each approach. The significance of RCCA in analyzing and interpreting the battery life cycle is emphasized, and other popular counting algorithms for predicting battery degradation in EVs and power grid applications are reviewed. In addition, improvement of RCCA for battery degradation assessment in the respective applications is reviewed and compared with the conventional approach. A comparison between the conventional and improved RCCA on an exemplar SOC data shows a significant low equivalent cycle at 2.1 cycles for the improved RCCA as compared to the compared to the conventional RCCA at 5.5 cycles. Furthermore, the equivalent life cycle can be updated in each time step using the improved RCCA rather than at SOC peaks or valleys in the conventional RCCA. The obtained equivalent cycle using RCCA can also be used to evaluate the aging process of other electronic equipment. However, RCCA has several limitations, including its sensitivity to test conditions such as negligible fluctuations in current and temperature. Despite these limitations, RCCA, together with open-source mechanisms and cloud data sharing, offers the opportunity to reform battery health assessment. This review serves as a useful reference for the design and operation of battery health diagnosis and prediction systems and provides guidance for future work on battery degradation assessment.

ACKNOWLEDGEMENTS

The authors acknowledge the Fundamental Research Grant Scheme (FRGS) with the grant number of FRGS/1/2021/TK0/UKM/02/9, funded by the Ministry of Higher Education (MOHE), Malaysia.

6. REFERENCES

[1] P. A. Owusu, S. Asumadu-Sarkodie, "A review of renewable energy sources, sustainability issues and climate change mitigation", Cogent Engineering, Vol. 3, No. 1, 2016, p. 1167990.

[2] M. Z. Daud, A. Mohamed, A. A. Ibrahim, M. A. Hannan, "Heuristic optimization of state-of-charge feedback controller parameters for output power dispatch of hybrid photovoltaic/battery energy storage system", Measurement, Vol. 49, 2014, pp. 15-25.

[3] L. Trahey *et al.* "Energy storage emerging: A perspective from the Joint Center for Energy Storage Research", Proceedings of the National Academy of Sciences, 2020, pp. 12550-12557.

[4] P. A. Dratsas, G. N. Psarros, S. A. Papathanassiou, "Battery energy storage contribution to system adequacy", Energies, Vol. 14, No. 16, 2021, p. 5146.

[5] M. A. Hannan *et al.* "Battery energy-storage system: A review of technologies, optimization objectives, constraints, approaches, and outstanding issues", Journal Energy Storage, Vol. 42, No. 7, 2021, p. 103023.

[6] W. L. Ai, H. Shareef, A. A. Ibrahim, A. Mohamed, "Optimal battery placement in photovoltaic based distributed generation using binary firefly algorithm for voltage rise mitigation", Proceedings of the IEEE International Conference on Power and Energy, Kuching, Malaysia, 1-3 December 2014, pp. 155-158.

[7] L. A. Wong, H. Shareef, A. Mohamed, A. A. Ibrahim, "Optimal Battery Sizing in Photovoltaic Based Distributed Generation Using Enhanced Opposition-Based Firefly Algorithm for Voltage Rise Mitigation", Scientific World Journal, Vol. 2014, No. 7, 2014, pp. 1-11.

[8] A. Maheshwari, N. G. Paterakis, M. Santarelli, M. Gibescu, "Optimizing the operation of energy storage using a non-linear lithium-ion battery degradation model", Applied Energy, Vol. 261, No. 12, 2020, p. 114360.

[9] K. C. Divya, J. Østergaard, "Battery energy storage technology for power systems-An overview", Electrical Power System Research, Vol. 79, No. 4, 2009, pp. 511-520.

[10] A. Barré, B. Deguilhem, S. Grolleau, M. Gérard, F. Suard, D. Riu, "A review on lithium-ion battery ageing mechanisms and estimations for automotive applications", Journal Power Sources, Vol. 241, 2013, pp. 680-689.

[11] J. Guo, Y. Li, K. Pedersen, D. Stroe, "Lithium-ion battery operation, degradation, and aging mechanism in Electric Vehicles: An Overview", Energies, Vol. 14, No. 5220, 2021, pp. 1-22.

[12] C. M. Tan, P. Singh, C. Chen, "Accurate real time online estimation of state-of-health and remaining useful life of Li ion batteries", Applied Sciences, Vol. 10, No. 21, 2020, pp. 1-16.

[13] L. Wu, X. Fu, Y. Guan, "Review of the Remaining Useful Life Prognostics of Vehicle Lithium-Ion Bat-

teries Using Data-Driven Methodologies", Applied Sciences, Vol. 6, No. 6, 2016, p. 166.

[14] K. A. Severson *et al.*, "Data-driven prediction of battery cycle life before capacity degradation", Nature Energy, Vol. 4, No. 5, 2019, pp. 383-391.

[15] R. Xiong, Y. Pan, W. Shen, H. Li, F. Sun, "Lithium-ion battery aging mechanisms and diagnosis method for automotive applications: Recent advances and perspectives", Renewable Sustainable Energy Review, Vol. 131, No. 5, 2020, p. 110048.

[16] X. Hu, L. Xu, X. Lin, M. Pecht, "Battery Lifetime Prognostics", Joule, Vol. 4, No. 2, 2020, pp. 310-346.

[17] M. S. Hosen, J. Jaguemont, J. Van Mierlo, M. Berecibar, "Battery lifetime prediction and performance assessment of different modeling approaches", iScience, Vol. 24, No. 2, 2021, p. 102060.

[18] S. Tamilselvi *et al.* "A review on battery modelling techniques", Sustainable, Vol. 13, No. 18, 2021, pp. 1-26.

[19] S. Jenu, A. Hentunen, J. Haavisto, M. Pihlatie, "State of health estimation of cycle aged large format lithium-ion cells based on partial charging", Journal Energy Storage, Vol. 46, No. 12, 2022, p. 103855.

[20] M. Dubarry, C. Truchot, B. Y. Liaw, "Synthesize battery degradation modes via a diagnostic and prognostic model", Journal Power Sources, Vol. 219, 2012, pp. 204-216.

[21] D. Widjajanto, B. Achsan, F. Rozaqi, A. Widyotriatmo, E. Leksono, "Estimasi Kondisi Muatan dan Kondisi Kesehatan Baterai VRLA dengan Metode RVP (Estimation of VRLA Batery's SOC and SOH Using SVR Method)", Jurnal Nasional Teknik Elektro dan Teknologi Informasi, Vol. 10, No. 2, 2021, pp. 178-187.

[22] T. Bak, S. Lee, "Accurate estimation of battery SOH and RUL based on a progressive lstm with a time compensated entropy index", Proceedings of the Annual Conference of the PHM Society, Scottsdale, AZ, USA, 23-26 September 2019, pp 1-10.

[23] A. Bocca, A. Sassone, D. Shin, A. Macii, E. Macii, M. Poncino, "An equation-based battery cycle life model for various battery chemistries", Proceedings of the IFIP/IEEE International Conference on Very Large Scale Integration, Daejeon, Korea, 5-7 October 2015, pp. 57-62.

[24] S. N. Motapon, E. Lachance, L. A. Dessaint, K. Al-Haddad, "A Generic Cycle Life Model for Lithium-Ion Batteries Based on Fatigue Theory and Equivalent Cycle Counting", IEEE Open Journal Industrial Electronics Society., Vol. 1, No. 8, 2020, pp. 207-217.

[25] T. K. Kim, S. C. Moon, "Novel Practical Life Cycle Prediction Method by Entropy Estimation of Li-Ion Battery", Electronics, Vol. 10, No. 4, 2021, p. 487.

[26] J. Adermann, D. Brecheisen, P. Wacker, M. Lienkamp, "Parameter Estimation of Traction Batteries by Energy and Charge Counting during Reference Cycles", Proceedings of the IEEE 86th Vehicular Technology Conference, Toronto, ON, Canada, 24-27 September 2017, pp. 1-7.

[27] W. Zhuo, A. V. Savkin, "Profit Maximizing Control of a Microgrid with Renewable Generation and BESS Based on a Battery Cycle Life Model and Energy Price Forecasting", Energies, Vol. 12, No. 15, 2019, p. 2904.

[28] N. Narayan, T. Papakosta, V. Vega-Garita, Z. Qin, J. Popovic-Gerber, P. Bauer, M. Zeman, "Estimating battery lifetimes in Solar Home System design using a practical modelling methodology", Applied Energy, Vol. 228, 2018, pp. 1629-1639.

[29] Y. L. Lee, T. Tjhung, "Rainflow Cycle Counting Techniques", Metal Fatigue Analysis Handbook, Elsevier, 2012, pp. 89-114.

[30] G. Marsh, C. Wignall, P. R. Thies, N. Barltrop, A. Incecik, V. Venugopal, L. Johanning, "Review and application of Rainflow residue processing techniques for accurate fatigue damage estimation", International Journal Fatigue, Vol. 82, 2016, pp. 757-765.

[31] V. Muenzel, J. de Hoog, M. Brazil, A. Vishwanath, S. Kalyanaraman, "A Multi-Factor Battery Cycle Life Prediction Methodology for Optimal Battery Management", Proceedings of the ACM Sixth International Conference on Future Energy Systems, Bangalore, India, 14-17 July 2015, pp. 57-66.

[32] B. Xu, A. Oudalov, A. Ulbig, G. Andersson, D. S. Kirschen, "Modeling of Lithium-Ion Battery Degradation for Cell Life Assessment", IEEE Transaction Smart Grid, Vol. 9, No. 2, 2018, pp. 1131-1140.

[33] Y. Shi, B. Xu, Y. Tan, B. Zhang, "A Convex Cycle-based Degradation Model for Battery Energy Storage Planning and Operation", Proceedings of the Annual American Control Conference, Milwaukee, WI, USA, 27-29 June 2017, pp. 4590-4596.

[34] Y. Shi, B. Xu, Y. Tan, D. Kirschen, B. Zhang, "Optimal Battery Control Under Cycle Aging Mechanisms in Pay for Performance Settings", IEEE Transaction Automatic Control, Vol. 64, No. 6, 2019, pp. 2324-2339.

[35] S. F. Schneider, P. Novak, T. Kober, "Rechargeable Batteries for Simultaneous Demand Peak Shaving and Price Arbitrage Business", IEEE Transaction Sustainable Energy, Vol. 12, No. 1, 2021, pp. 148-157.

[36] D. Rosewater, A. Headley, F. A. Mier, S. Santoso, "Optimal Control of a Battery Energy Storage System with a Charge-Temperature-Health Model", Proceedings of the IEEE Power & Energy Society General Meeting, August 2019, pp. 1-5.

[37] P. Singh, S. Dhundhara, Y. P. Verma, N. Tayal, "Optimal battery utilization for energy management and load scheduling in smart residence under demand response scheme", Sustainable Energy, Grids Networks, Vol. 26, 2021, p. 100432.

[38] B. Foggo, N. Yu, "Improved Battery Storage Valuation Through Degradation Reduction", IEEE Transaction Smart Grid, Vol. 9, No. 6, 2018, pp. 5721-5732.

[39] J. O. Lee, Y. S. Kim, "Novel battery degradation cost formulation for optimal scheduling of battery energy storage systems", International Journal Electrical Power Energy System, Vol. 137, No. 10, 2022, p. 107795.

[40] A. Soleimani, V. Vahidinasab, "Integrating Battery Condition and Aging into the Energy Management of an Active Distribution Network", Proceedings of the Smart Grid Conference, Tehran, Iran, 18-19 December 2019, pp. 1-6.

[41] A. Soleimani, V. Vahidinasab, J. Aghaei, "A Linear Stochastic Formulation for Distribution Energy Management Systems Considering Lifetime Extension of Battery Storage Devices", IEEE Access, Vol. 10, 2022, pp. 44564-44576.

[42] C. Tang, J. Xu, Y. Sun, S. Liao, F. Zhang, L. Ma, "Stochastic battery energy storage scheduling considering cell degradation and distributed energy resources", International Transaction Electrical Energy System, Vol. 29, No. 7, 2019, pp. 1-20.

[43] M. Chawla, R. Naik, R. Burra, H. Wiegman, "Utility energy storage life degradation estimation method", Proceedings of the IEEE Conference on Innovative Technologies for an Efficient and Reliable Electricity Supply, Waltham, MA, USA, 27-29 September 2010, pp. 302-308.

[44] K. Abdulla *et al.* "Optimal Operation of Energy Storage Systems Considering Forecasts and Battery Degradation", IEEE Transaction Smart Grid, Vol. 9, No. 3, 2018, pp. 2086-2096.

[45] J. Martins, S. Spataru, D. Sera, D.-I. Stroe, A. Lashab, "Comparative Study of Ramp-Rate Control Algorithms for PV with Energy Storage Systems", Energies, Vol. 12, No. 7, 2019, p. 1342.

[46] H. Beltran, J. Barahona, R. Vidal, J. C. Alfonso, C. Arino, E. Perez, "Ageing of different types of batteries when enabling a PV power plant to enter electricity markets", Proceedings of the 42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, Italy, 23-26 October 2016, pp. 1986-1991.

[47] M. J. E. Alam, T. K. Saha, "Cycle-life degradation assessment of Battery Energy Storage Systems caused by solar PV variability", Proceedings of the IEEE Power and Energy Society General Meeting, Boston, MA, USA, 17-21 July 2016, pp. 1-5.

[48] M. A. Hossain, M. R. Mahmud, H. R. Pota, "Optimal Energy Scheduling of Residential Building with Battery Cost", Proceedings of the 9th International Conference on Power and Energy Systems, Perth, WA, Australia, 10-12 December 2019, pp. 1-6.

[49] L. Ochoa-Eguilegor, N. Goitia-Zabaleta, A. Gonzalez-Garrido, A. Saez-de-Ibarra, H. Gaztanaga, A. Hernandez, "Optimized Market Bidding of Energy Storage Systems for Dynamic Containment Service", Proceedings of the IEEE International Conference on Environment and Electrical Engineering and IEEE Industrial and Commercial Power Systems Europe, Bari, Italy, 7-10 September 2021, pp. 1-6.

[50] G. Karmiris, T. Tengner, "Control method evaluation for battery energy storage system utilized in renewable smoothing", Proceedings of the 39th Annual Conference of the IEEE Industrial Electronics Society, Vienna, Austria, 10-13 November 2013, pp. 1566-1570.

[51] A. Bouakkaz, A. J. Gil Mena, S. Haddad, M. L. Ferrari, "Scheduling of Energy Consumption in Standalone Energy Systems Considering the Battery Life Cycle", Proceedings of the IEEE International Conference on Environment and Electrical Engineering and IEEE Industrial and Commercial Power Systems Europe, Madrid, Spain, 9-12 June 2020, pp. 1-4.

[52] T. Dragicevic, H. Pandzic, D. Skrlec, I. Kuzle, J. M. Guerrero, D. S. Kirschen, "Capacity Optimization of Renewable Energy Sources and Battery Storage in an Autonomous Telecommunication Facility", IEEE Transaction Sustainable Energy, Vol. 5, No. 4, 2014, pp. 1367-1378.

[53] Y. R. Lee, H. J. Kim, M. K. Kim, "Optimal Operation Scheduling Considering Cycle Aging of Battery Energy Storage Systems on Stochastic Unit Commitments in Microgrids" Energies, Vol. 15, No. 6, 2022, p. 2107.

[54] O. Boqtob, H. El Moussaoui, H. El Markhi, T. Lamhamdi, "Energy Scheduling of Isolated Microgrid with Battery Degradation Cost using Hybrid Particle Swarm Optimization with Sine Cosine Acceleration Coefficients", International Journal Renewable Energy Research, Vol. 10, No. v10i2, 2020, pp. 704-715.

[55] C. Lyu, Y. Jia, Z. Xu, M. Shi, "Real-Time Operation optimization of Islanded Microgrid with Battery Energy Storage System", Proceedings of the IEEE Power & Energy Society General Meeting, Montreal, QC, Canada, 2-6 August 2020, pp. 1-5.

[56] G. He, Q. Chen, C. Kang, P. Pinson, Q. Xia, "Optimal Bidding Strategy of Battery Storage in Power Markets Considering Performance-Based Regulation and Battery Cycle Life", IEEE Transaction Smart Grid, Vol. 7, No. 5, 2016, pp. 2359-2367.

[57] C. A. Correa-Florez, A. Gerossier, A. Michiorri, G. Kariniotakis, "Stochastic operation of home energy management systems including battery cycling", Applied Energy, Vol. 225, 2018, pp. 1205-1218.

[58] J. Li, A. M. Gee, M. Zhang, W. Yuan, "Analysis of battery lifetime extension in a SMES-battery hybrid energy storage system using a novel battery lifetime model", Energy, Vol. 86, 2015, pp. 175-185.

[59] C. Pan, S. Tao, H. Fan, M. Shu, Y. Zhang, Y. Sun, "Multi-Objective Optimization of a Battery-Supercapacitor Hybrid Energy Storage System Based on the Concept of Cyber-Physical System", Electronics, Vol. 10, No. 15, 2021, p. 1801.

[60] B. Lian, D. Yu, C. Wang, S. Le Blond, R. W. Dunn, "Investigation of energy storage and open cycle gas turbine for load frequency regulation", Proceedings of the 49th International Universities Power Engineering Conference, Cluj-Napoca, Romania, 2-5 September 2014, pp. 1-6.

[61] S. Loew, A. Anand, A. Szabo, "Economic model predictive control of Li-ion battery cyclic aging via online rainflow-analysis", Energy Storage, Vol. 3, No. 3, 2021, pp. 1-23.

[62] A. Anand, S. Loew, C. L. Bottasso, "Economic control of hybrid energy systems composed of wind turbine and battery", Proceedings of the European Control Conference, Delft, Netherlands, 29 June - 2 July 2021, pp. 2565-2572.

[63] A. Z. Obaro, J. L. Munda, A. A. YUSUFF, "Modelling and Energy Management of an Off-Grid Distributed Energy System: A Typical Community Scenario in South Africa", Energies, Vol. 16, No. 2, 2023, p. 693.

[64] A. Singh, P. Pareek, L. P. M. I. Sampath, L. Goel, H. B. Gooi, H. D. Nguyen, "A Stress-Cognizant Optimal Battery Dispatch Framework for Multimarket Participation", IEEE Transaction Industrial Informatics, Vol. 20, No. 5, 2024, pp. 7259-7268.

[65] C. Lorf, R. F. Martinez-Botas, N. Brandon, "26,500km Down the Pan-American Highway in an Electric Vehicle A Battery's Perspective", SAE International Journal Alternative Powertrains, Vol. 1, No. 1, 2012, p. 2012-01-0123.

[66] S. Li, H. He, J. Li, P. Yin, H. Wang, "Machine learning algorithm based battery modeling and management method: A Cyber-Physical System perspective", Proceedings of the 3rd Conference on Vehicle Control and Intelligence, Hefei, China, 21-22 September 2019, pp. 1-4.

[67] S. Li, H. He, C. Su, P. Zhao, "Data driven battery modeling and management method with aging phenomenon considered", Applied Energy, Vol. 275, No. 3, 2020, p. 115340.

[68] Y. Bai, H. He, J. Li, S. Li, Y. Wang, Q. Yang, "Battery anti-aging control for a plug-in hybrid electric vehicle with a hierarchical optimization energy management strategy", Journal Cleaner Production, Vol. 237, 2019, p. 117841.

[69] J. J. Espinosa, S. Ruiz, "Optimal vehicle-to-grid strategy for a fleet of EVs considering the batteries aging", Engineer e Investigation, Vol. 39, No. 2, 2019, pp. 69-75.

[70] M. Sandelic, A. Sangwongwanich, F. Blaabjerg, "Incremental Degradation Estimation Method for Online Assessment of Battery Operation Cost", IEEE Transaction Power Electronics, Vol. 37, No. 10, 2022, pp. 11497-11501.

[71] B. Vatandoust, A. Ahmadian, M. A. Golkar, A. Elkamel, A. Almansoori, M. Ghaljehei, "Risk-Averse Optimal Bidding of Electric Vehicles and Energy Storage Aggregator in Day-Ahead Frequency Regulation Market", IEEE Transaction Power System, Vol. 34, No. 3, 2019, pp. 2036-2047.

[72] S. Li, J. Li, C. Su, Q. Yang, "Optimization of Bi-Directional V2G Behavior With Active Battery Anti-Aging Scheduling", IEEE Access, Vol. 8, 2020, pp. 11186-11196.

[73] B. Verbrugge *et al.*, "Reliability Assessment of SiC-Based Depot Charging Infrastructure with Smart and Bidirectional (V2X) Charging Strategies for Electric Buses", Energies, Vol. 16, No. 1, 2022, p. 153.

[74] D. Pelosi, M. Longo, D. Zaninelli L. Barelli, "Experimental Investigation of Fast–Charging Effect on Aging of Electric Vehicle Li–Ion Batteries", Energies, Vol. 16, No. 18, 2023, p. 6673.

[75] J. Huang, S. Wang, W. Xu, C. Fernandez, Y. Fan, X. Chen, "An Improved Rainflow Algorithm Combined with Linear Criterion for the Accurate Li-ion Battery Residual Life Prediction", International Journal Electrochemical Science, Vol. 16, No. 7, 2021, p. 21075.

[76] J. Huang, S. Wang, W. Xu, W. Shi, C. Fernandez, "A Novel Autoregressive Rainflow—Integrated Moving Average Modeling Method for the Accurate State of Health Prediction of Lithium-Ion Batteries", Processes, Vol. 9, No. 5, 2021, p. 795.

[77] B. Gundogdu, D. T. Gladwin, "A Fast Battery Cycle Counting Method for Grid-Tied Battery Energy Storage System Subjected to Microcycles", Proceedings of the International Electrical Engineering Congress, Krabi, Thailand, 7-9 March 2018, pp. 1-4.

[78] R. Diao, Z. Hu, Y. Song, "Subgradient of Cycle-Based Aging Cost Function and Its Application in Optimal Operation of Battery Energy Storage System With Multiple Subsystems", IEEE Transaction Energy Conversion, Vol. 39, No. 1, 2024, pp. 625-643.

In Memoriam
Prof. Goran Martinović, PhD

(August 7th, 1969, Orahovica – October 16th, 2025, Osijek)

With deep sorrow and respect, we remember our dear friend and colleague, professor Goran Martinović, PhD, a distinguished researcher and university professor who left us forever on October 16th, 2025. Professor Goran Martinović, PhD, was born in Orahovica in 1969. After completing his master's degree in 2000, he obtained his doctorate in 2004 in the scientific field of Technical Sciences, area of Computing, at the Faculty of Electrical Engineering and Computing, University of Zagreb. Following his doctorate, he worked as an assistant professor from December 2004, as an associate professor from 2009, as a full professor from 2012, and since 2017 as a full professor with tenure at FERIT Osijek, where he also served as Vice Dean for Science and International Cooperation. During his scientific and teaching career, he participated in establishing numerous undergraduate, graduate, and postgraduate study programs at FERIT.

Together with his collaborators and friends, Professor Martinović worked in an environment marked by professionalism, closeness, and humanity. He inspired everyone around him with his hard work and optimism, always accompanied by a cheerful spirit and enthusiasm. Over nearly three decades at FERIT, through his vision and dedication, he guided many young people toward scientific and professional excellence for the common good. Many former FERIT students will remember him as an outstanding expert, lecturer, and professor. His passing leaves a great void in the academic community, as well as in the hearts of colleagues, collaborators, and students who had the privilege of learning and working alongside him. His exceptional contributions will remain permanently inscribed in our academic and professional community, and his dedication and humanity will be deeply missed.

In his teaching work, he designed and delivered numerous courses and helped develop educational materials. He supervised more than 200 defended graduate and undergraduate theses, two master's theses, and ten doctoral dissertations. His scientific, teaching, and professional activities spanned embedded computer systems, software engineering, distributed computer systems, real-time systems, computer intelligence, and data analysis.

His outstanding contribution to the development of science, industry, and technology is reflected in his extensive involvement in numerous national and international projects, notably DATACROSS, TEAMSOC21, HKO, PROGRESS, and many others. He actively participated in two COST actions and collaborated with numerous foreign and domestic institutions.

He authored or co-authored more than 250 scientific and professional papers, cited in over 1500 other publications, as well as seven edited books. He organized more than twenty professional workshops within the European EuroCC and EuroCC 2 projects, further contributing to the development and dissemination of knowledge in computing. As the founder and Editor-in-Chief of the international scientific journal International Journal of Electrical and Computer Engineering Systems, and as a member of editorial and program boards of numerous journals and conferences, he left an indelible mark on the academic community. He was a member of KoREMA, IEEE, ACM, and the Croatian Academy of Engineering (HATZ). For his contributions to engineering education and his work within IEEE, he received numerous recognitions, including the IEEE Outstanding SMC Chapter Award (2019), the IEEE Croatia Section Award (2008), and the IEEE Letter of Appreciation (2006). He was also the recipient of the annual "Rikard Podhorsky" Award of the Croatian Academy of Engineering.

In 2009, the Faculty of Electrical Engineering, Computer Science and Information Technology Osijek launched the international scientific-professional journal International Journal of Electrical and Computer Engineering Systems, with Prof. Martinović serving as its founding Editor-in-Chief. The first issue was published in June 2010. Since then, it has been published regularly, initially in two issues per year, and now in ten issues per year. Prof. Dr. sc. Martinović performed the role of editor-in-chief with dedication and exceptional expertise from the very beginning until his death, making an invaluable contribution to the development and reputation of the journal. His success is best evidenced by the fact that the International Journal of Electrical and Computer Engineering Systems is included in reputable databases, such as Scopus and Web of Science.

Irena Galić, Mario Vranješ
Faculty of Electrical Engineering,
Computer Science and Information Technology Osijek

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia.

About this Journal

The International Journal of Electrical and Computer Engineering Systems publishes original research in the form of full papers, case studies, reviews and surveys. It covers theory and application of electrical and computer engineering, synergy of computer systems and computational methods with electrical and electronic systems, as well as interdisciplinary research.

Topics of interest include, but are not limited to:

- Power systems
- Renewable electricity production
- Power electronics
- Electrical drives
- Industrial electronics
- Communication systems
- Advanced modulation techniques
- RFID devices and systems
- Signal and data processing
- Image processing
- Multimedia systems
- Microelectronics
- Instrumentation and measurement
- Control systems
- Robotics
- Modeling and simulation
- Modern computer architectures
- Computer networks
- Embedded systems
- High-performance computing
- Parallel and distributed computer systems
- Human-computer systems
- Intelligent systems
- Multi-agent and holonic systems
- Real-time systems
- Software engineering
- Internet and web applications and systems
- Applications of computer systems in engineering and related disciplines
- Mathematical models of engineering systems
- Engineering management
- Engineering education

Paper Submission

Authors are invited to submit original, unpublished research papers that are not being considered by another journal or any other publisher. Manuscripts must be submitted in doc, docx, rtf or pdf format, and limited to 30 one-column double-spaced pages. All figures and tables must be cited and placed in the body of the paper. Provide contact information of all authors and designate the corresponding author who should submit the manuscript to https://ijeces.ferit.hr. The corresponding author is responsible for ensuring that the article's publication has been approved by all coauthors and by the institutions of the authors if required. All enquiries concerning the publication of accepted papers should be sent to ijeces@ferit.hr.

The following information should be included in the submission:

- paper title;
- full name of each author;
- full institutional mailing addresses;
- e-mail addresses of each author;
- abstract (should be self-contained and not exceed 150 words). Introduction should have no subheadings;
- manuscript should contain one to five alphabetically ordered keywords;
- all abbreviations used in the manuscript should be explained by first appearance;
- all acknowledgments should be included at the end of the paper:
- authors are responsible for ensuring that the information in each reference is complete and accurate. All references must be numbered consecutively and citations of references in text should be identified using numbers in square brackets. All references should be cited within the text;
- each figure should be integrated in the text and cited in a consecutive order. Upon acceptance of the paper, each figure should be of high quality in one of the following formats: EPS, WMF, BMP and TIFF;
- corrected proofs must be returned to the publisher within 7 days of receipt.

Peer Review

All manuscripts are subject to peer review and must meet academic standards. Submissions will be first considered by an editor-in-chief and if not rejected right away, then they will be reviewed by anonymous reviewers. The submitting author will be asked to provide the names of 5 proposed reviewers including their e-mail addresses. The proposed reviewers should be in the research field of the manuscript. They should not be affiliated to the same institution of the manuscript author(s) and should not have had any collaboration with any of the authors during the last 3 years.

Author Benefits

The corresponding author will be provided with a .pdf file of the article or alternatively one hardcopy of the journal free of charge.

Units of Measurement

Units of measurement should be presented simply and concisely using System International (SI) units.

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003

Published: semiannually

Copyright

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.

Postal Address

Faculty of Electrical Engineering,
Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b
31000 Osijek, Croatia

IJECES Copyright Transfer Form

(Please, read this carefully)

This form is intended for all accepted material submitted to the IJECES journal and must accompany any such material before publication.

TITLE OF ARTICLE (hereinafter referred to as "the Work"):

COMPLETE LIST OF AUTHORS:

The undersigned hereby assigns to the IJECES all rights under copyright that may exist in and to the above Work, and any revised or expanded works submitted to the IJECES by the undersigned based on the Work. The undersigned hereby warrants that the Work is original and that he/she is the author of the complete Work and all incorporated parts of the Work. Otherwise he/she warrants that necessary permissions have been obtained for those parts of works originating from other authors or publishers.

Authors retain all proprietary rights in any process or procedure described in the Work. Authors may reproduce or authorize others to reproduce the Work or derivative works for the author's personal use or for company use, provided that the source and the IJECES copyright notice are indicated, the copies are not used in any way that implies IJECES endorsement of a product or service of any author, and the copies themselves are not offered for sale. In the case of a Work performed under a special government contract or grant, the IJECES recognizes that the government has royalty-free permission to reproduce all or portions of the Work, and to authorize others to do so, for official government purposes only, if the contract/grant so requires. For all uses not covered previously, authors must ask for permission from the IJECES to reproduce or authorize the reproduction of the Work or material extracted from the Work. Although authors are permitted to re-use all or portions of the Work in other works, this excludes granting third-party requests for reprinting, republishing, or other types of re-use. The IJECES must handle all such third-party requests. The IJECES distributes its publication by various means and media. It also abstracts and may translate its publications, and articles contained therein, for inclusion in various collections, databases and other publications. The IJECES publisher requires that the consent of the first-named author be sought as a condition to granting reprint or republication rights to others or for permitting use of a Work for promotion or marketing purposes. If you are employed and prepared the Work on a subject within the scope of your employment, the copyright in the Work belongs to your employer as a work-for-hire. In that case, the IJECES publisher assumes that when you sign this Form, you are authorized to do so by your employer and that your employer has consented to the transfer of copyright, to the representation and warranty of publication rights, and to all other terms and conditions of this Form. If such authorization and consent has not been given to you, an authorized representative of your employer should sign this Form as the Author.

Authors of IJECES journal articles and other material must ensure that their Work meets originality, authorship, author responsibilities and author misconduct requirements. It is the responsibility of the authors, not the IJECES publisher, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it.

- The undersigned represents that he/she has the authority to make and execute this assignment.
- For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
- The undersigned agrees to indemnify and hold harmless the IJECES publisher from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

_____ _____

Author/Authorized Agent **Date**

CONTACT
International Journal of Electrical and Computer Engineering Systems (IJECES)
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Josip Juraj Strossmayer University of Osijek
Kneza Trpimira 2b
31000 Osijek, Croatia
Phone: +38531224600,
Fax: +38531224605,
e-mail: ijeces@ferit.hr