Indonesian Journal of Electrical Engineering and Computer Science
Vol. 40, No. 2, November 2025, pp. 1100~1108
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v40.i2.pp1100-1108

Enhancing document text classification using hybrid deep contextual and correlation network

Shilpa, Shridevi Soma
Department of Computer Science and Engineering, PDA College of Engineering, Visvesvaraya Technological University, Belagavi, India

Article Info

Article history:

Received Jul 27, 2024
Revised Aug 28, 2025
Accepted Oct 15, 2025

Keywords:

Deep learning
HDCCNet
Natural language processing
Text classification
Text representation

ABSTRACT

Document analysis involves the extraction and processing of information from documents, a task increasingly automated through the use of deep learning (DL) technologies. Despite the high predictive power of DL models, their black-box nature poses challenges to transparency and interpretability, hindering their integration into the industry. This paper introduces the hybrid deep contextual and correlation network (HDCCNet), a novel methodology designed to improve both the accuracy and interpretability of multi-category classification tasks. HDCCNet leverages a hybrid layer category correlation module to deepen category connections, thereby enhancing the understanding and prediction of category interrelations. To address potential prediction divergence, residual connections are incorporated, ensuring stable and reliable performance. Furthermore, HDCCNet reduces model parameters, accelerating convergence and making the model more efficient. This efficiency is particularly beneficial for practical applications, allowing faster deployment and scalability. By bridging the gap between DL's capabilities and industry needs for transparency, HDCCNet provides a robust solution for automated document processing, paving the way for broader adoption of DL technologies in business environments.

Corresponding Author:

Shilpa
Department of Computer Science and Engineering, PDA College of Engineering
Visvesvaraya Technological University
Belagavi, India
Email: shilpa_122023@rediffmail.com

1. INTRODUCTION

Deep learning (DL) has made major advancements in document analysis over the past few years [1]. This method offers a lot of potential for automating administrative tasks related to processing documents. In prediction tasks, deep neural networks have proven to perform well. However, one significant barrier to the safe integration of these technologies into corporate processes continues to be a lack of transparency. Unstructured data was frequently used to represent text data in the early days of digital computing. Technological advances in text data storage have led to a major growth in the field of information retrieval (IR). Thanks to advances in technology, text classification and textual data processing may now be done more automatically [2]. Experts concluded at the beginning of the project that mathematical indexing was insufficient to attain higher levels of accuracy. As the benefits of allowing computers to comprehend and interpret human language become more apparent, natural language processing (NLP), is becoming more and more popular.

There are several noteworthy similarities across various document collections. When the destination domain is similar to or equal to the source domain, transfer learning is the process of moving knowledge or information from one domain to another [3]. Using transfer learning techniques occasionally requires specific features to exist in both the source and destination domains. It might be difficult to train transfer learning algorithms for the target domain classifier, especially if the target texts lack categories. One problem is using terminology specific to the destination domain that does not exist in the source domain. Documents are frequently categorized into one or more groups within a high-dimensional, sparse space during the classification process. Numerous statistical and artificial intelligence methods have been used to solve the problem of document classification [4].

The four primary stages of document categorization are pre-processing, document prediction, classifier training, and feature extraction. Duplicate data are removed at the pre-processing stage of data analysis. Words that are uncommon in this specific context are evaluated as part of the review process. N-gram models are frequently used in a wide range of applications for feature extraction. One of the models used most frequently in NLP is the bag-of-words model. When user-generated information is enclosed in square brackets and followed by a number, it is regarded as a source citation. This setup has been installed. Terms that are regularly used have a property called term frequency (TF), which is the basic concept underlying the bag-of-words approach. Document classification is based on the ideas of term frequency and inverse document frequency (TF-IDF), which combines TF and IDF with IDF [5]. The TensorFlow (TF) framework is a fundamental component of our work. A text classifier’s capacity to rapidly increase readability through word frequency is one of its key advantages. By concentrating on the most crucial sections of the document, the classifier’s study may locate the keywords associated with the content.

However, it’s crucial to remember that books can include a lot of technical jargon. The combination of the bag-of-words architecture and the integrated frequency-based features yields a sparse and vast feature space. Algorithmic feature selection can assist in efficiently and rapidly locating a certain collection of pertinent qualities [6]. To build models, the support vector machines (SVM) machine learning technique makes use of certain attributes. The genetic programming (GP) approach allows the computer to autonomously choose a set of characteristics while building classifiers, in contrast to models that rely on preselected features [7]. Moreover, converting words into numerical vectors is a keystone of DL methodologies. Classifiers should be trained using the conversion procedure.

Document segmentation solutions by dividing lengthy texts into smaller, more digestible phrases or chunks by using slider windows or sequential cutting algorithms. When arranged hierarchically, phrases or sections appear before the entire representation of the text. Saifullah *et al.* [8] used a technique in their study that involved breaking up the raw data into smaller pieces. The units were then analyzed using the BERT basic model. For classification, this model makes use of one recurrent layer or one extra transformer. Once this feature is put into practice, all of the outputs may be shared, which makes use of an interactive transformer model, and improved sentence modeling even more. Sentence representation accuracy is improved since this method considers the context of the full document. The existing approaches, which assume hierarchical representations for long texts, often fail to capture the interactions across several feature levels and struggle to communicate information at the same level. However, the current algorithms have not fully assessed the structural information included in lengthy texts.

DL has revolutionized document analysis, offering significant potential to automate document processing in business workflows. Despite their predictive power, the black-box nature of DL models hinders safe integration in the industry due to a lack of transparency and interpretability. Businesses require not only accurate models but also ones that can be understood and trusted by users, making interpretability a critical factor in adopting DL solutions. This paper addresses the challenge of enhancing interpretability while maintaining high performance in document classification tasks. In the realm of document classification, traditional methods have relied heavily on manual feature engineering and simple machine learning algorithms, which, although interpretable, often lack the predictive power of modern DL approaches. Recent advancements in DL have demonstrated superior performance by automatically extracting complex features from raw text data. However, these models’ decisions are often opaque, making it difficult to diagnose errors or understand the model’s reasoning process. This lack of transparency can lead to resistance to adopting these technologies in sensitive applications where understanding the decision-making process is crucial.

- Enhanced category correlation learning: our methodology introduces a hybrid deep contextual and correlation network (HDCCNET), which significantly deepens the connections between categories, leading to improved understanding and prediction of category interrelations. This enhances the overall accuracy and reliability of multi-category classification tasks.
- Improved interpretability: by incorporating residual connections and focusing on essential features, our approach not only prevents divergence in predictions but also offers greater transparency and interpretability. This allows users to understand the model’s decision-making process, addressing a critical need for trustworthy DL models in the industry.

- Efficiency and scalability: our method reduces the number of model parameters compared to traditional approaches, accelerating convergence and making the model more efficient. This efficiency is particularly beneficial for practical applications, enabling faster deployment and scalability in various business workflows.

2. RELATED WORK

The field of document analysis has seen substantial advancements with the integration of DL techniques, enhancing the efficiency and accuracy of document classification tasks. Early approaches relied heavily on manual feature engineering and traditional machine learning algorithms, such as logistic regression (LR) and SVMs, which, while interpretable, often lacked the predictive power of DL models. Modern methodologies leverage deep neural networks, particularly convolutional neural networks (CNNs) and graph convolutional networks (GCNs), which have demonstrated superior performance by automatically extracting complex features from raw text data. Notable works include the use of hierarchical transformers and BERT-based models, which have shown significant improvements in handling long documents and capturing intricate text representations. Despite these advancements, a persistent challenge remains the black-box nature of DL models, which hinders their interpretability and thus their acceptance in the industry. Recent studies have focused on enhancing model transparency through techniques like attention mechanisms and explainable AI frameworks, aiming to bridge the gap between high performance and the need for model interpretability. Saifullah *et al.* [9] reported a property collection that was done by hand. The articles were then categorized using the previously mentioned attributes as a guide. The ratios of textual to non-textual portions, column layouts, content density, and font sizes in comparison were among the many factors considered throughout the selection process.

The data was categorized using a decision tree that was trained using the provided attributes, it discussed the concepts of document similarity and a querying technique designed specifically for document image databases. The structural similarity that is geometrically invariant is found. The following statement is a response to previous research conducted by relevant parties. Pujar *et al.* [10] employed AdaBoost in combination with an ensemble of K-means clustering-based classifiers to find articles for their study quickly. The identification process was based on the low-level binary image pixel density data processing. A novel method for automatically recognizing picture anchor templates from document images. The templates may be used for a variety of tasks, such as data extraction and document categorization. In their inquiry [11] proposed utilizing codebooks as a means of assessing document picture similarity. The document is recursively broken into smaller bits using this technique. The attributes mentioned before were utilized to retrieve documents from the database that shared the same characteristics. The authors enhanced their previous findings by training and retrieving document photos that belong to the same category using an unsupervised random forest classifier. The achievement was attained by optimizing the calculated representations.

In the subsequent year [12], the authors achieved state-of-the-art performance on the tax form and table retrieval tasks using the same techniques. The performance evaluation’s underlying assumption was that insufficient training data provided a detailed account of the first application of deep CNN in the field of document picture categorization. This approach outperformed previous hand-coded feature engineering methodologies in terms of speed. In their research publication introduced the DeepDoc Classifier as an example of the potential applications of transfer learning [13]. The study makes use of a deep CNN with the AlexNet architecture. The weights of the network were initialized using a pre-trained model that was trained on the large ImageNet dataset. The dataset consists of 1.28 million training pictures that are categorized into 1000 distinct groups. The performance of earlier approaches was significantly improved by transforming the original convolutional layers into flexible feature extractors. To increase the accuracy of the classification findings, the textual data from a commercial optical character recognition (OCR) system with the raw image data [14]. After it was recovered, an NLP model translated the text into the feature space. Employed an extreme learning machine to manipulate frozen convolutional layers that were trained using an AlexNet model. The group was able to increase output without compromising accuracy requirements.

Pappagari *et al.* [1] evaluated the performance of visual geometry group (VGG), ResNet, and GoogLeNet in classifying document pictures using benchmark datasets. The study finds that significant gains may be achieved with pretrained image classification networks, which have been trained on a large amount of categorized data. They presented a two-stream network that could generate output based on both textual and visual inputs. In contrast to previous studies, an algorithm was employed to evaluate the textual stream’s features and pinpoint the most significant ones. They combined OCR predictions with image data using a methodology akin to previous work. However, the performance was enhanced by employing bidirectional encoder representations from transformers (BERT) as the NLP framework, which looked at the viability of applying self-supervised representation learning for document image classification. However, the study conducted by the researchers was limited to a few old-fashioned self-supervised activities, such as Jigsaw puzzles. The study effectively demonstrated the limitations of the strategy in terms of generating meaningful representations.

3. PROPOSED METHOD

This model proposes a HDCCNet model composed of two primary modules: the $DCRINet$ and the $DCNet$ module. The $DCRINet$ module enhances document representation by incorporating category information, treating category information as positive samples of document information through contrastive learning. This process allows for an in-depth exploration of the relationships between documents and their corresponding categories. BERT is employed as a feature extractor to obtain semantic features from both the documents and the categories, which are then processed by the $DCRINet$ module. The output from the $DCRINet$ module serves as the input for the $DCNet$ module. In $DCNet$, the original category predictions are further refined using correlation knowledge by training multiple weight matrices. This module focuses on identifying relevant combinations of original category predictions to enhance the overall prediction accuracy. By integrating category semantic information and category correlation, HDCCNET achieves precise multi-category classification results. Figure 1 shows the proposed HDCCN architecture.

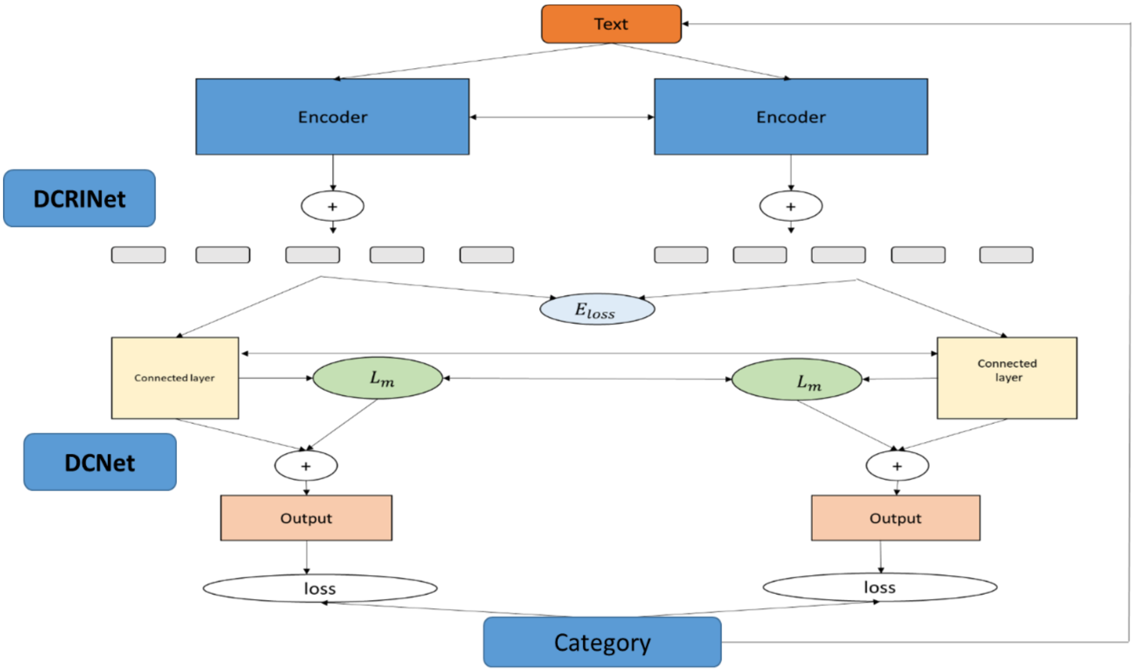

Figure 1. Proposed HDCCN architecture

3.1. Initial analysis

Assume the dataset $=\{(z_k, a_k)\}_{k=1}^{P}$, here z_k is the original document and the $k-th$ text as, $z_k = \{y_1, y_2, \dots, y_V\}$, V represents the length of the input document y_k is the $k-th$ word of the document. $a_k \in \{0,1\}^N$ denotes the adjacent category set for z_k, N is the category set of the dataset and P is the total number of examples in the dataset, the classifier here computes the probability $r_k$for each category being true where $r_k = \{r_1, r_2, \dots, r_N\}$, the loss between r_k and a_k is given by (1). Before being given as input to the BERT model, the document z_k is considered, considering a model with p layers the hidden representation is denoted by the $p-th$ layer as given in (2).

$$Loss(r_k, a_k) = -\frac{1}{N}\sum_{n\in N} \ [a_n log r_n + (1-a_n) log(1-r_n)] \quad (1)$$

$$\alpha_{bert}^{(p)}(z_k) = \{j_{cls}^{(p)}, j_1^{(p)}, \dots \dots., j_V^{(p)}\} \quad (2)$$

3.2. Deep contextual representation integrator (DCRI) Net

As the semantic characteristic of the document and category, we make use of the embedded layer in enhanced BERT. Nevertheless, investigating the connections between the semantic characteristics of the text and the category requires more than just using the layer of the final layer. In the training phase, the document features and category features are synchronously extracted, and the text feature is denoted as given in (3). To address the aforementioned issue, we thus suggest a document-category contrastive learning loss function. the encoding of M document features utilizing the features of M as multi-category sets related to the document hence the number of samples is 2M, K={1, …, 2M}. The index of the $K-th$ document as category (k) and the negative sample remains as in K. The loss function of the expression is shown in (4).

$$j = concat(j_{cls}^{(-1)}, j_{cls}^{(-2)}, j_{cls}^{(-3)}, j_{cls}^{(-4)}, j_{cls}^{(-5)}) \quad (3)$$

$$N_{Enc} = \sum_{k=1}^{2m} \; -log\frac{exp(j_k \cdot \frac{j_{label(k)}}{\vartheta})}{\sum_{m\in K/k} exp(j_k \cdot \frac{j_m}{\vartheta})} \quad (4)$$

Here ϑ is the temperature co-efficient that can distinguish between positive and negative samples. j_k is the concatenated vector of the layers obtained from j to z_k. Two main factors contribute to Category $DCRINet$ success. Using the semantic information of the category is the first step in creating a document representation with category information guidance. Using this data, a language model called BERT is instructed to extract characteristics from the document information that are pertinent to the category. It also incorporates an organized framework for connecting the semantic data. In this technique, the category feature encoding is used as a positive sample for the category and the document contents.

3.3. Deep correlation Net (DCNet)

The two parts of $DCNet$ are the raw category predictions and the L_m computational unit, which converts the raw category predictions into improved category predictions depending on category correlations. The $DCNet$ model is defined by the given (5). Here A and A are the output and input of the $DCNet$ model, here a is the raw category prediction before $DCNet$ model and A is the category prediction model with correlation represented by the L_m.

$$A = a + L_m(a) \quad (5)$$

The simplest design for a category correlation module adds a linear layer after y, similar to Cornet, for correlated category prediction. However, this approach yields shallow category correlations. To deepen the connections, we propose a multi-layer category correlation module where each layer's output serves as the next layer's input, improving category correlations. To prevent divergence in predictions, we incorporate residual connections using the original category predictions. This design, which has fewer parameters than the total number of categories, focuses on learning deep correlations and accelerates model convergence. Here x is the input for the correlation layer and β is the activation function. Category prediction as the output of $info$ achieves high accuracy in guiding the category correlation prediction as input of $DCNet$. The category prediction is the input for $DCNet$. The effect is amplified through residual connections.

$$clayer(x) = \beta(Yx + d) \quad (6)$$

3.4. HDCCNET model training

The end-to-end based multi-category classification model is made up of $DCRINet$ and $DCNet$ model, the goal is to reduce the target loss as N_s that consists of N_{Enc}, $Loss(r_k^{text}, a_k)$ and $Loss(r_k^{label}, a_k)$, this is defined as given in (7). Here α is the coefficient associated with the contrast learning of the categorized document between the losses, r_k^{text} is the last probability of the $k-th$ document semantic info as input. r_k^{label} is the last probability of semantic info of the category set relevant with the $k-th$ document as input.

$$N_s = \alpha N_{Enc} + \; Loss(r_k^{text}, a_k) + Loss(r_k^{label}, a_k) \quad (7)$$

4. PERFORMANCE EVALUATION

The performance evaluation demonstrates that the proposed system (PS) method significantly outperforms other state-of-the-art techniques in document classification on the exAAPD2 and exPFD datasets, respectively. This notable performance gap highlights the PS method's advanced techniques and optimizations, making it a robust and reliable choice for high-accuracy classification tasks. The clustering of other methods' scores within narrower ranges suggests their comparable effectiveness, yet PS's superior results underscore its potential advantages in practical applications.

4.1. Dataset

- exAAPD2: the extended annotated automatic phoneme data (exAAPD) [15] dataset is a comprehensive resource designed for advanced speech processing research. It features high-quality audio recordings with detailed, time-aligned phoneme annotations, supporting a wide array of applications including speech and phoneme recognition, speaker identification, and emotion analysis. The dataset encompasses multiple languages, providing a valuable tool for cross-linguistic studies and comparisons. Each recording comes with extensive metadata covering speaker demographics, recording conditions, and linguistic details, enhancing its research utility. The diversity of speakers, accents, and environments represented in the dataset ensures its applicability across various contexts and research goals.
- exPFD: the extended phonetic frame data (exPFD) dataset is a sophisticated resource designed for phonetic and speech processing research. It includes high-fidelity audio recordings with detailed, time-aligned phonetic frame annotations, providing precise category for phonetic events. The dataset supports multiple languages and dialects, making it an invaluable tool for comparative phonetic studies. It features extended annotations that encompass phonetic context, speaker attributes, and environmental conditions, alongside comprehensive phonetic transcriptions. Rich metadata accompanies each recording, detailing speaker demographics, recording setups, and linguistic backgrounds. This extensive dataset is ideal for applications in phonetic research, speech processing, and language studies, enabling in-depth analyses of phonetic properties and variations across diverse linguistic contexts. Researchers must adhere to ethical guidelines to ensure the responsible use and integrity of the data.

4.2. State-of-art techniques

- LR, KimCNN [16], FastText [17], XML-CNN [18], HAN [19], and regularized LSTM [20] are examples of conventional document classification techniques. Hedwig5 is a DL toolkit that offers pre-implemented document classification models. It provides various approaches for document classification.
- The LDC has developed the hierarchical graph convolutional network (HGCN), a model capable of processing both the micro level of a word graph and the macro level of a section graph. Several document categorization techniques based on GCN have been developed, including Text GCN [21] and BertGCN [22]. The focus of these techniques is to classify nodes on graphs at the corpus level. In addition, there are other techniques available for categorizing nodes on document-level networks, including TL GNN [23] and HyperGAT [24].
- The transformer-based LDC techniques come in several forms. Among these is BERT+TextRank 15], a truncation baseline extension of BERT. It entails using TextRank to get an additional set of 512 tokens. Another method that can be used instead of the BERT+TextRank model is BERT+Random [25]. It chooses up to 512 tokens at random from texts. Furthermore, RoBERT [1] is an extension of BERT that permits its use in LDC. On top of the segment representations, RoBERT applies a recurrence layer after segmentation. To BERT [1], A transformer layer is suggested as a substitute for the recurrent layer in the RoBERT model. To manage lengthy texts, a Transformer model variation called the Longformer [26] combines local and global attention processes. The BigBird [27] is an additional Longformer extension that adds random attention. In addition, a hierarchical model called the Hi-Transformer [3] concentrates on obtaining a document's whole context to model sentences [28].

4.3. Results

The Figure 2 chart displays the Test.F1 scores for various classification methods, revealing that most methods perform similarly, with scores ranging from approximately 66 to 68. Methods such as KimCNN [16], TL-GNN [22], XML-CNN [18], FastText [17], TextGCN [21], LSTM(Reg) [20], HAN [19], LR, BertGCN [24], BERT+Random [15], HGCN+BERT, Hi-Transformer [15], BERT+Textrank [25], RoBERTa [1], Longformer [27], HyperGAT [23], BigBird [3], HGCN-BigBird [15], ToBERT [1], HMGCN+BERT [28], and HMGCN+BigBird [28] all fall within this range, indicating comparable effectiveness in the task at hand. However, the PS method stands out significantly, achieving a Test.F1 score of approximately 79, which is noticeably higher than the others. This suggests that the PS method incorporates more advanced techniques or optimizations, leading to its superior performance. The clear gap between PS and the other methods highlights its robustness and potential advantages, making it a promising choice for tasks requiring high accuracy and reliability in classification.

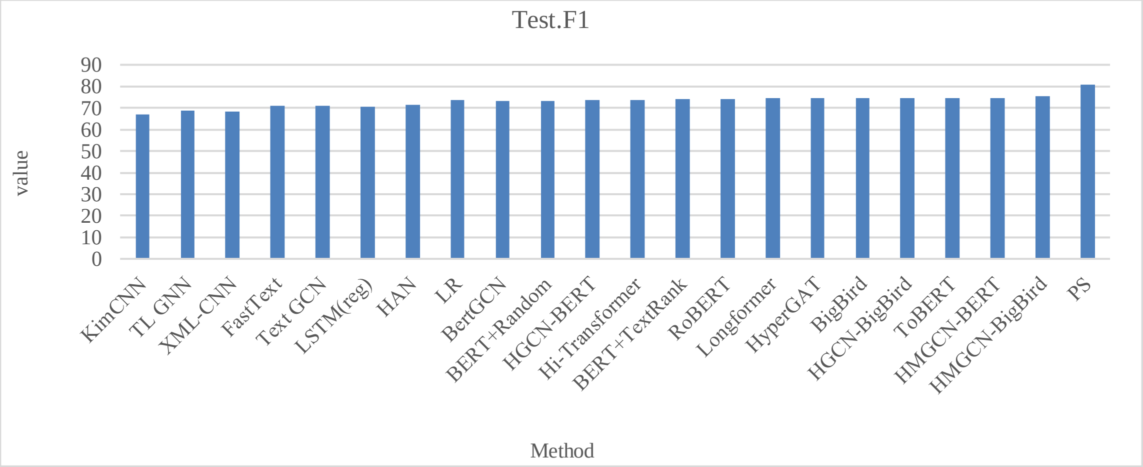

Figure 2. Test.F1 results for exAAPD2 dataset in comparison with the existing state-of-art techniques with PS

Figure 3 depicts the Test.F1 scores for various classification methods for exPFD dataset, highlighting their performance differences. The methods range from TL-GNN, which scores around 76, to PS, which achieves the highest score at approximately 88. Most methods, including BertGCN [24], TextGCN [21], LR, HyperGAT [23], KimCNN [16], FastText [17], XML-CNN [18], Hi-Transformer, HAN, LSTM(Reg), Longformer, BigBird, BERT+Random, RoBERTa, HGCN-BERT, HGAT-BigBird, ToBERT, HGCN+BigBird, BERT+Textrank, HMGCN+BERT, and HMGCN+BigBird, fall within a narrower range of 78 to 84. This indicates a close clustering of performance among these methods, suggesting they offer similar effectiveness for the given task. However, the PS method distinctly outperforms all others with its Test.F1 score of 88, indicating a significant performance advantage. This suggests that the PS method utilizes more advanced or optimized techniques, setting it apart from the other methods and making it a particularly robust choice for achieving high accuracy in classification tasks.

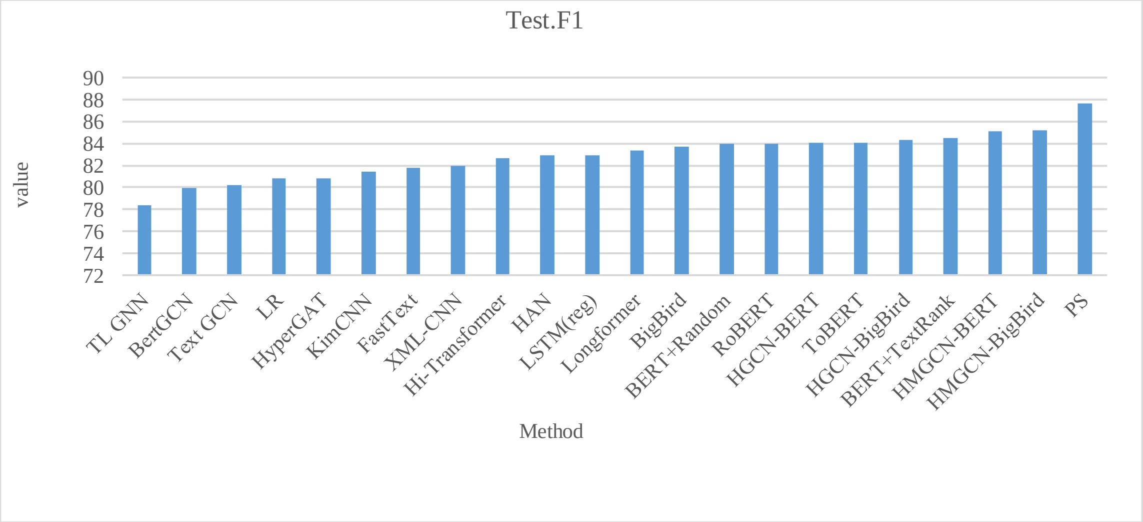

Figure 3. Test.F1 result for exPFD dataset in comparison with the existing state-of-art techniques with PS

5. CONCLUSION

In conclusion, the HDCCNET presents a significant advancement in the field of document analysis by addressing the dual challenges of accuracy and interpretability in multi-category classification tasks. By introducing a multi-layer category correlation module, HDCCNET enhances the depth of category connections, leading to improved prediction accuracy and a more nuanced understanding of category interrelations. The incorporation of residual connections ensures stability and prevents divergence in predictions, contributing to the model's reliability. Furthermore, HDCCNET's design, which reduces the number of parameters, not only accelerates model convergence but also makes it more efficient and scalable for practical applications. This efficiency is particularly valuable in business environments where computational resources and time are often limited. By bridging the gap between the powerful capabilities of DL and the practical need for transparency and interpretability, HDCCNET provides a robust and scalable solution for automated document processing. The contributions of HDCCNET pave the way for the broader adoption of DL technologies in various business workflows, offering a reliable, transparent, and efficient approach to document analysis. Future work will explore the application of HDCCNET to other domains and further enhancements to improve its robustness and applicability in diverse real-world scenarios.

ACKNOWLEDGEMENT

I would like to express our sincere gratitude to all those who have supported and contributed to this research project. Primarily, I extend our heartfelt thanks to our guide for her unwavering guidance, invaluable insights, and encouragement throughout the research process.

FUNDING INFORMATION

The authors state no funding involved.

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] R. Pappagari, P. Zelasko, J. Villalba, Y. Carmiel, and N. Dehak, "Hierarchical transformers for long document classification," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019, pp. 838–844, doi: 10.1109/ASRU46091.2019.9003958.

[2] X. Zhang, F. Wei, and M. Zhou, "HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5059–5069, doi: 10.18653/v1/P19-1499.

[3] C. Wu, F. Wu, T. Qi, and Y. Huang, "Hi-transformer: hierarchical interactive transformer for efficient and effective long document modeling," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 2021, vol. 2, pp. 848–853, doi: 10.18653/v1/2021.acl-short.107.

[4] Saifullah, S. Agne, A. Dengel, and S. Ahmed, "The reality of high performing deep learning models: a case study on document image classification," *IEEE Access*, vol. 12, pp. 103537–103564, 2024, doi: 10.1109/ACCESS.2024.3425910.

[5] S. Saifullah, S. Agne, A. Dengel, and S. Ahmed, "DocXclassifier: towards a robust and interpretable deep neural network for document image classification," *International Journal on Document Analysis and Recognition*, vol. 27, no. 3, pp. 447–473, Sep. 2024, doi: 10.1007/s10032-024-00483-w.

[6] R. Powalski, Ł. Borchmann, D. Jurkiewicz, T. Dwojak, M. Pietruszka, and G. Pałka, "Going full-TILT boogie on document understanding with text-image-layout transformer," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12822 LNCS, 2021, pp. 732–747.

[7] J. Ferrando *et al.*, "Improving accuracy and speeding up document image classification through parallel systems," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12138 LNCS, 2020, pp. 387–400.

[8] Saifullah, S. A. Siddiqui, S. Agne, A. Dengel, and S. Ahmed, "Are deep models robust against real distortions? A case study on document image classification," in *2022 26th International Conference on Pattern Recognition (ICPR)*, Aug. 2022, vol. 2022-Augus, pp. 1628–1635, doi: 10.1109/ICPR56361.2022.9956167.

[9] S. Saifullah, S. Agne, A. Dengel, and S. Ahmed, "Analyzing the potential of active learning for document image classification," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 26, no. 3, pp. 187–209, Sep. 2023, doi: 10.1007/s10032-023-00429-8.

[10] P. Pujar, A. Kumar, and V. Kumar, "Efficient plant leaf detection through machine learning approach based on corn leaf image classification," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 13, no. 1, p. 1139, Mar. 2024, doi: 10.11591/ijai.v13.i1.pp1139-1148.

[11] S. H. Sreedhara, V. Kumar, and S. Salma, "Efficient big data clustering using adhoc fuzzy C means and auto-encoder CNN," in *Lecture Notes in Networks and Systems*, vol. 563, 2023, pp. 353–368.

[12] R. Khasawneh and R. Kornreich, "Explaining data-driven document classifications," *MIS Quarterly: Management Information Systems*, vol. 3, no. 4, pp. 781–791, 2014.

[13] T. Vermeire, D. Brughmans, S. Goethals, R. M. B. de Oliveira, and D. Martens, “Explainable image classification with evidence counterfactual,” *Pattern Analysis and Applications*, vol. 25, no. 2, pp. 315–335, May 2022, doi: 10.1007/s10044-021-01055-y.

[14] O. Lang *et al.*, “Explaining in style: training a GAN to explain a classifier in StyleSpace,” in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 673–682, doi: 10.1109/ICCV48922.2021.00073.

[15] T. Liu, Y. Hu, B. Wang, Y. Sun, J. Gao, and B. Yin, “Hierarchical graph convolutional networks for structured long document classification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 10, pp. 8071–8085, Oct. 2023, doi: 10.1109/TNNLS.2022.3185295.

[16] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751, doi: 10.3115/v1/D14-1181.

[17] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, vol. 2, pp. 427–431, doi: 10.18653/v1/E17-2068.

[18] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, “Deep learning for extreme multi-label text classification,” in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 2017, pp. 115–124, doi: 10.1145/3077136.3080834.

[19] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1480–1489, doi: 10.18653/v1/N16-1174.

[20] A. Adhikari, A. Ram, R. Tang, and J. Lin, “Rethinking complex neural network architectures for document classification,” in *Proceedings of the 2019 Conference of the North*, 2019, vol. 1, pp. 4046–4051, doi: 10.18653/v1/N19-1408.

[21] L. Yao, C. Mao, and Y. Luo, “Graph convolutional networks for text classification,” *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, pp. 7370–7377, 2019, doi: 10.48550/arXiv.1809.05679.

[22] L. Huang, D. Ma, S. Li, X. Zhang, and H. Wang, “Text level graph neural network for text classification,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 3442–3448, doi: 10.18653/v1/D19-1345.

[23] K. Ding, J. Wang, J. Li, D. Li, and H. Liu, “Be more with less: hypergraph attention networks for inductive text classification,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4927–4936, doi: 10.18653/v1/2020.emnlp-main.399.

[24] Y. Lin *et al.*, “BertGCN: transductive text classification by combining GNN and BERT,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 1456–1462, doi: 10.18653/v1/2021.findings-acl.126.

[25] H. Park, Y. Vyas, and K. Shah, “Efficient classification of long documents using transformers,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022, vol. 2, pp. 702–709, doi: 10.18653/v1/2022.acl-short.79.

[26] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: the long-document transformer,” *arXiv*, 2020, [Online]. Available: http://arxiv.org/abs/2004.05150.

[27] M. Zaheer *et al.*, “Big bird: transformers for longer sequences,” *Advances in Neural Information Processing Systems*, vol. 2020-Decem, pp. 17283–17297, Jan. 2021, [Online]. Available: http://arxiv.org/abs/2007.14062.

[28] T. Liu, Y. Hu, J. Gao, Y. Sun, and B. Yin, “Hierarchical multi-granularity interaction graph convolutional network for long document classification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1762–1775, 2024, doi: 10.1109/TASLP.2024.3369530.

BIOGRAPHIES OF AUTHOR

Mrs. Shilpa received her bachelors degree in Computer Science and Engineering from the Visvesvaraya Technological University, BELGAUM - India in 2010 and Master Degree in Computer Science and Engineering from same University in 2012. She is currently pursuing her Ph.D. degree from the same university. She is presently working as assistant professor in Computer Science and Engineering Dept. Sharnbasva University Kalaburagi, Karnataka, – India. Her primary area of interest is image processing, machine learning, and pattern recognition. She can be contacted at email: shilpa_122023@rediffmail.com.

Dr. Shridevi Soma working presently as professor in Department of Computer Science and Engineering, Poojya Doddappa Appa College of Engineering, Kalaburagi. She has 18 years of Teaching and 10 years of Research Experience, and completed her B.E., M.Tech., and Ph.D. in Computer Science and Engineering. Her research area includes digital image processing and pattern recognition, cloud computing, internet of things, big data analytics. She published more than 30 Research papers in above mentioned areas, also Guiding Research Students. She has also received grant for establishment of “Cloud Computing Lab” from VGST. She can be contacted at email: shridevisoma@gmail.com.