ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

Enhancing Healthcare Data Classification with CNN

[1] Md Sohaib Iqubal, [2] Dr. Rajeev Kumar

[1] [2] Dept of Computer Science and Engineering National Institute of Technology Hamirpur, Himachal Pradesh, India
Corresponding Author Email: [1] 195524@nith.ac.in, [2] rajeev@nith.ac.in

Abstract— This paper presents a novel methodology aimed at refining the classification of Electronic Health Record (EHR) text, focusing on overcoming challenges posed by unstructured narratives and complex abbreviations. Our approach emphasizes improving predictive precision in patient diagnosis by strategi- cally organizing critical elements such as Symptoms, History, and Medications within the EHR framework. Leveraging Convolu- tional Neural Networks (CNNs) and the GloVE pretrained model, our study aims to discern intricate features in medical summaries, potentially reshaping the landscape of healthcare data classifi- cation. By incorporating insights from recent advancements in health informatics, including the utilization of Health Record Summaries and the integration of deep learning techniques, our methodology seeks to bolster the efficiency and efficacy of patient care practices. Through keyword extraction from the database, and model architecture design, we strive to enhance deep learning models capabilities in healthcare tasks such as clustering patients based on common features. Additionally, our exploration of automated medical record labeling using the MT Samples Dataset underscores the potential of deep learning in addressing the challenges of handling vast amounts of healthcare data. By integrating rule-based, machine learning, and hybrid approaches, and emphasizing feature selection for optimal section recognition, our methodology offers a comprehensive framework for advancing clinical predictive analytics and processing volu- minous health-related data. Overall, this paper contributes to the ongoing efforts in leveraging cutting-edge technologies to improve healthcare data management and patient care outcomes.

Index Terms— Deep learning (DL), Convolutional Neural Net- work(CNN), Natural Language Processing(NLP),Word Embed- ding.

I. INTRODUCTION

This research explores integrating advanced deep learning methodologies into Electronic Health Record (EHR) classification in healthcare data management. By leveraging Convolutional Neural Networks (CNNs), the study aims to enhance healthcare analytics and decision-making in health informatics. It focuses on the transformative potential of EHR systems in redefining healthcare data organization and utilization. Over time, the utilization of Health Record Summaries has increased, containing various patient information. However, organizing this data efficiently requires discerning crucial de- tails amidst structured and unstructured information, including specialized medical notes, which poses challenges for machine learning algorithms due to errors and abbreviations.

Traditional methodologies for health record analysis, such as logistic regression and support vector machines (SVM), are being replaced by deep learning techniques. These methods construct intricate features capable of capturing complex data dependencies, offering superior performance with minimal pre-processing time. This paper specifically explores auto- mated medical record labeling using the MT Samples Dataset, aiming to enhance deep learning models' capabilities in health- care tasks such as anomaly detection in medical images and clustering patients based on common features.

Deep learning's advantage lies in its minimal programmer intervention and autonomous decision-making. Unlike traditional machine learning, deep learning autonomously manages validation and normalization functions, reducing preprocessing time. Deep learning excels at learning optimal features directly from data, eliminating the need for manual feature crafting, which is time-consuming and complex.

Artificial neural networks (ANNs) play a crucial role in deep learning, serving as the foundation for various methodologies. These networks, comprised of interconnected nodes, autonomously glean insights from data, uncovering complex relationships that traditional methods may miss.

The proposed model involves dataset preparation, keyword extraction, and text cleaning. The model architecture includes layers tailored to enhance text data analysis and improve keyword categorization effectiveness.

Regarding classification, rule-based, machine learning (ML), and hybrid approaches are employed, with an emphasis on feature selection to optimize section recognition methods. However, comparing performance across studies remains chal- lenging due to the absence of standardized evaluation methods. As medical records continue to grow exponentially, auto- mated systems leveraging Natural Language Processing (NLP) and deep learning are essential for efficient organization and categorization. These advancements facilitate rapid diagnosis and interventions, making automated document classification crucial in advancing clinical predictive analytics and processing voluminous health-related data.

II. RELATED WORK

In the domain of health-related text classification, specialized methods have emerged as effective solutions. Supervised topic models, such as prediction-focused supervised Latent Dirichlet Allocation (LDA) introduced by Jason et al. [1], leverage clinical data to estimate hidden factors predicting output variables, enhancing coherence of topics while ensur- ing efficient prediction. Notably, utilizing the Merck Man- ual dataset, studies have employed word-level matrices and word2vec embeddings, exhibiting superior accuracy compared to alternative methods [2]. Convolutional neural networks (CNNs) have also proven effective in categorizing clinical narrative surgery data, demonstrating substantial improvement in F1-Score.

Frameworks like Med Cat, suggested by Fodeh et al. [3], offer innovative approaches to structuring medical notes, particularly in contexts such as Post Traumatic Stress Disorder. These frameworks utilize manual annotation and Natural Language Processing (NLP) techniques to extract detailed features, transforming them into specialized concept-category hierarchies for improved understanding and organization of medical information.

Design principles emphasizing hierarchical representation learning have shown promise in enhancing the performance of text classification models. Very deep convolutional networks and deep pyramid convolutional neural networks leverage deep stacks of local operations to achieve superior accuracy by learning high-level hierarchical representations of sentences. Furthermore, systems like MedEx by Xu et al. [4] demonstrate efficacy in extracting medication information from clinical nar- ratives, contributing significantly to healthcare safety, quality, and clinical research.

In processing natural language to derive essential concepts from clinical narratives, Clinical Named Entity Recognition (NER) is crucial. Wu et al. [5] demonstrated the superiority of deep neural network architectures, achieving state-of-the-art results in clinical NER tasks. Deep learning approaches, including CNNs and graph convolutional networks, have also shown superiority in phenotyping tasks, outperforming concept-based extraction methods and yielding higher F1 scores in classifying medical relationships from clinical narra- tives. These findings collectively underscore the effectiveness of deep learning approaches in extracting and classifying information from clinical narratives, paving the way for ad- vancements in healthcare data analysis and decision-making. Additionally, recent research by Lee et al. [6] presents a comparative study on clinical named entity recognition meth- ods using Korean clinical texts. They compared conventional named entity recognition (NER) methods, namely dictionary- lookup-based string matching and conditional random fields (CRFs), on the clinical texts of rheumatism patients in South Korea. Their study demonstrates that CRFs outperform string matching in extracting most semantic types from clinical notes, with a median F1 score of 0.761. The results indicate that CRFs are a promising candidate for implementing clinical NER in Korean clinical narrative documents.

Advancements in Natural Language Processing (NLP) have contributed significantly to Electronic Health Records (EHRs) for computational phenotyping, as reviewed by Smith et al. [7]. Their study highlights the various applications of NLP- based computational phenotyping, including diagnosis catego- rization, novel phenotype discovery, clinical trial screening, pharmacogenomics, drug-drug interaction (DDI), and adverse drug event (ADE) detection, as well as genome-wide and phenome-wide association studies. They discuss the progress made in algorithm development and resource construction for computational phenotyping, with supervised machine learning models being favored for their ability to acquire both classification patterns and structures from data. Moreover, deep learning and unsupervised learning methods have gained attention for their performance and ability to find novel phenotypes. The integration of heterogeneous data sources has also shown promise in improving model performance. Despite these advancements, challenges such as model interpretability, generalizability, and proper characterization of feature rela- tions in clinical narratives remain.

Recent research by Chen et al. [8] explores the impact of pretrained language models on negation and speculation detection in cross-lingual medical text. They introduce a system for cross-lingual and domain-independent negation and speculation detection, particularly focusing on the biomedical scientific literature and clinical narrative. Their study considers negation and speculation detection as a sequence-labeling task, proposing approaches using bidirectional long short- term memory (Bi-LSTM) and conditional random field, as well as bidirectional encoder representations for transformers (BERT) with fine-tuning for named entity recognition (NER). Evaluation on English and Spanish languages using biomedical and review text datasets yielded promising results, with F- measures ranging from 80.8% to 91.7%.

Research by Xie et al. [9] provides a systematic review of challenges and methodologies in utilizing deep learning for temporal data representation in electronic health records (EHRs). They identify major challenges including data irregularity, heterogeneity, sparsity, and model opacity in temporal EHR data, and evaluate novel methodologies for addressing them through deep learning solutions. Their study underscores the potential of deep learning in addressing challenges associ- ated with temporal EHR data, highlighting the importance of incorporating clinical domain knowledge and enhancing model interpretability for effective clinical prediction modeling and data utilization.

Moreover, Rajkomar et al. [10] propose a representation of patients' entire raw Electronic Health Records (EHR) records based on the Fast Healthcare Interoperability Re- sources (FHIR) format. They demonstrate that deep learning methods using this representation are capable of accurately predicting multiple medical events without site-specific data harmonization. Their approach achieved high accuracy for tasks such as predicting in-hospital mortality, 30-day unplanned readmission, prolonged length of stay, and all of a patient's final discharge diagnoses, outperforming traditional predictive models. This scalable and accurate approach holds promise for personalized medicine and improving healthcare quality.

Ford et al. [11] conducted a systematic review examining the extraction of information from the free text of Electronic Medical Records (EMRs) to improve case detection. They found that incorporating information from text into case-detection algorithms significantly improved algorithm sensitivity and area under the receiver operating characteristic compared to using coded parts of EMRs alone. Their review underscores the importance of utilizing both structured codes and unstructured text in EMRs for accurate case detection, thereby improving research quality in health-related studies.

Wang and Shah [12] developed a novel machine learning algorithm, the 'Semi-supervised Set Covering Machine' (S3CM), to extract diagnoses and investigation results from unstructured free text in Electronic Health Records (EHRs). Their algorithm achieved high recall and precision in detecting coronary angiogram results and ovarian cancer diagnoses, out- performing other algorithms tested. This approach highlights the potential of semi-supervised machine learning techniques in automatically identifying relevant information from unstruc- tured text in EHRs, thereby improving efficiency and accuracy in medical research.

Leevy et al. [13] conducted a survey on recurrent neural network (RNN) and conditional random field (CRF) models for de-identification of medical free text in Electronic Health Records (EHRs). Their work revealed that RNN models, particularly long short-term memory (LSTM) algorithms, gen- erally outperformed CRF models and other systems, such as rule-based algorithms. Additionally, they found that hybrid or ensemble systems containing joint LSTM-CRF models showed no advantage over individual LSTM and CRF models. Their findings highlight the importance of considering model perfor- mance, overfitting issues, and diversity during experimentation when developing de-identification systems for medical free text.

Moreover, Miotto et al. [14] provide a comprehensive review of the recent literature on applying deep learning technologies to advance the health care domain. They suggest that deep learning approaches could be the vehicle for translating big biomedical data into improved human health. However, they also note limitations and needs for improved methods development and applications, especially in terms of ease-of- understanding for domain experts and citizen scientists. Their work highlights the potential of deep learning in healthcare while emphasizing the importance of developing interpretable architectures to bridge deep learning models and human inter- pretability.

III. METHODOLOGY

Given the sensitive nature of patient identity and the confidentiality surrounding medical transcriptions—comprising dialogues between healthcare professionals and patients, coupled with assessments of their prevailing medical conditions—access to such data is restricted, thereby presenting challenges for studies and model development focused on text classification. In this section, we delineate the methodology employed in constructing a predictive model aimed at catego- rizing medical specialty domains utilizing keyword extraction from transcriptions. [15] We expound on the approach used for dataset acquisition, taking into consideration the intricacies of data privacy and ethical considerations. Furthermore, we elu- cidate the performance evaluation metrics employed to gauge the efficacy of the model, encompassing training accuracy, validation accuracy, F1 score, precision, and recall. These metrics offer valuable insights into the model's proficiency in accurately classifying medical specialty categories based on extracted transcription keywords.

Also, it's crucial to emphasize the significance of ensuring compliance with data protection regulations and ethical guidelines. This entails anonymizing data to safeguard patient privacy and uphold ethical standards in research practices. [16] Moreover, considering the dynamic nature of medical termi- nology and practices, continuous refinement and validation of the model against diverse datasets could enhance its robustness and generalizability across different healthcare settings. [17]

A. Dataset

The dataset utilized in this study was meticulously curated from the extensive MT samples database, encompassing a vast array of 30,000 medical transcription samples across a diverse spectrum of 39 distinct medical specialties, spanning from dentistry to surgery. Each entry within this rich repository is meticulously annotated with pertinent details, including sam- ple name, description, and a comprehensive set of keywords intricately associated with the corresponding medical specialty. Such meticulous curation ensures the richness and depth of the dataset, enabling a multifaceted exploration of doctor- patient interactions and diagnostic evaluations across a myriad of medical domains.

This expansive and meticulously assembled dataset stands as an invaluable asset for healthcare research and model development endeavors, providing a robust foundation for con- ducting in-depth analyses and formulating predictive models. By encompassing a broad spectrum of medical specialties, the dataset facilitates a nuanced understanding of the intricacies inherent in various medical disciplines, thereby enabling the development of more accurate and specialized predictive mod- els.

Moreover, the open accessibility of this dataset plays a pivotal role in fostering academic exploration and collaboration within the research community. By removing barriers such as licensing restrictions, we can afford the freedom to delve into the intricacies of medical text classification, thereby facilitating the dissemination of knowledge and fostering advancements in the field. This open-access approach not only promotes transparency but also encourages innovation by enabling researchers to build upon existing work and explore novel methodologies in medical text analysis.

B. GloVe Word Vector

GloVe, which stands for Global Vectors for Word Representation, represents a groundbreaking leap in machine learning methodology, introduced by Jeffrey Pennington, Richard Socher, and Christopher D. Manning in 2014 [18]. This innovative algorithm operates by tapping into global word-to- word relationships derived from a given word corpus, thereby constructing vector representations for individual words. These resulting vectors hold immense utility as they can be effec- tively utilized to train on new data, facilitating predictive tasks and enhancing natural language understanding.

One of the key strengths of GloVe lies in its provision of word vectors across various dimensions, including 50D, 100D, 200D, and 300D, offering flexibility and granularity in representation. This versatility allows for nuanced analysis and interpretation of textual data across a wide spectrum of applications. Moreover, GloVe is meticulously designed to

learn word vectors in a manner that ensures the dot product of vectors closely mirrors the logarithm of the probability of word co-occurrence, thus rendering the resulting vectors adept at discerning word analogies and semantic relationships.

In this paper, GloVe's 100-dimensional embeddings from the Glove 6B are employed, indicating the utilization of GloVe vectors with 100 dimensions. Furthermore, supplementary hyperparameters are incorporated to facilitate the generation of out-of-vocabulary word embeddings, which involves inte- grating n-gram embeddings into the process. This meticulous approach not only enhances the coverage and representation of words but also accommodates those absent from the original GloVe vocabulary, thereby enriching the overall effectiveness of the model and improving its performance in real-world scenarios.

GloVe represents a sophisticated and versatile tool in the realm of natural language processing, offering researchers and practitioners powerful capabilities for text analysis, semantic understanding, and predictive modeling. Its robustness, flex- ibility, and effectiveness make it a cornerstone in modern machine learning research and applications.

C. Convolutional Neural Network

Convolutional Neural Network

In this study, our primary aim is to harness the power of Deep Learning techniques for effectively categorizing medical records into their respective domains. Central to our approach is the integration of pre-trained GloVe embeddings, specif- ically utilizing the "glove.6b.100d" model derived from the Wikipedia-2014 and Gigaword-5 datasets.

Convolutional Neural Networks (CNNs) have emerged as a staple in image classification tasks due to their layered archi- tecture comprising convolutional layers, nonlinear activation functions, and a mix of convolution, max-pooling, and dense layers. While traditionally associated with image processing, CNNs exhibit remarkable potential in Natural Language Pro- cessing (NLP) tasks, particularly text classification [19]. In NLP, words are represented as vectors or word embeddings, capturing semantic meanings. When applied to CNNs for NLP tasks, text undergoes convolution similar to images, albeit with adjustments in filter sizes to accommodate text features.

Prior studies have underscored the effectiveness of CNNs in computer vision tasks, leveraging their ability to capture compositional structures and spatial invariance. Despite the inherent disparities between image and text data, CNNs offer computational efficiency, particularly with GPU acceleration. This computational prowess is invaluable for processing extensive textual data efficiently. In our proposed model, input sequences consist of keywords extracted from the dataset, which undergo a preprocessing phase to ensure uniformity in representation through punctuation removal and lowercase conversion. These keywords are then tokenized and partitioned into training and testing datasets, with a portion reserved for validation.

Next, the pre-trained GloVe embeddings are integrated into the model's embedding layer, transforming the input keywords into 1000-dimensional vectors. These embeddings feed into a

1D convolutional layer, as illustrated in Figure 1, depicting the model's architectural flow. The text undergoes conversion into sequences of integer IDs, forming an embedding matrix that maps pre-trained embeddings to indexed words. This embedding matrix serves as the foundation for the Keras Embedding layer, upon which the 1D convolutional layer is constructed. [20]

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 11 Issue 7 July 2024

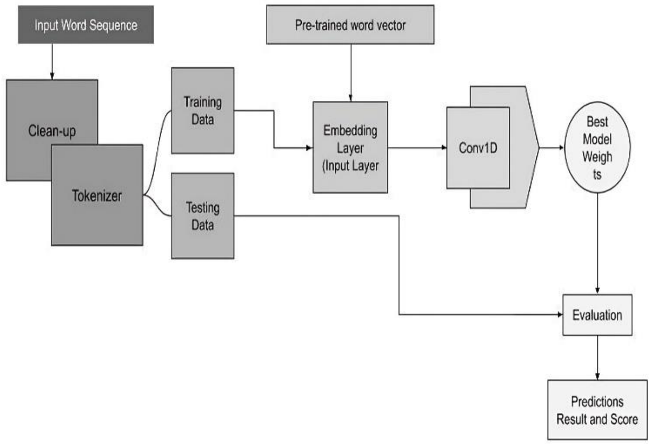


Fig. 1. Model Flow-chart

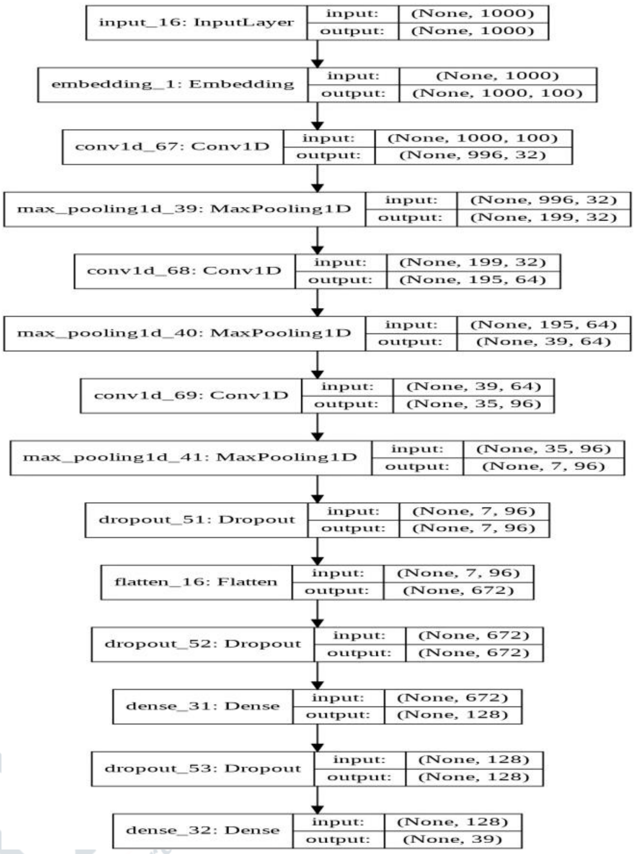
An additional convolutional layer Figure 2, is introduced to accommodate lengthy input sequences, followed by a max-pooling layer to emphasize crucial text segments. To counter overfitting, a dropout layer randomly deactivates connections during training iterations. Finally, a flatten layer reshapes inputs from the dropout layer, preparing them for the dense layer, ensuring effective feature learning and representation for accurate categorization of health record data.

During testing, complete discharge data are inputted, and keyword extraction is performed, followed by GloVe or TF-IDF filtering to remove common words. This database by MTsamples, already contain pre-extracted keywords, stream-lining integration into the model. Furthermore, the model demonstrates adaptability in classifying subspecialties within medical domains, enhancing its versatility across diverse test- ing scenarios and datasets.

The utilization of Convolutional Neural Networks in medical text classification underscores their efficacy in capturing intricate features and patterns within healthcare narratives, paving the way for enhanced categorization accuracy and improved clinical decision support systems.

IV. EXPERIMENTAL ANALYSIS AND RESULT

The CNN model was applied to the MT Samples dataset, with multiple layers iteratively added and removed to optimize efficiency. Throughout the training process, the training accuracy consistently improved with each epoch, reaching a peak value of 0.87. Similarly, the validation accuracy exhibited an upward trend, culminating at 0.89, as illustrated in Figure 3. During the validation phase, the model demonstrated com- mendable performance metrics, with a recall score of 0.888, precision of 0.98, and an F1 score of 0.939. These metrics collectively underscore the model's efficacy in accurately classifying medical transcription samples.



input_16: InputLayer	input:	(None, 1000)
	output:	(None, 1000)
embedding_1: Embedding	input:	(None, 1000)
	output:	(None, 1000, 100)
conv1d_67: Conv1D	input:	(None, 1000, 100)
	output:	(None, 996, 32)
max_pooling1d_39: MaxPooling1D	input:	(None, 996, 32)
	output:	(None, 199, 32)
conv1d_68: Conv1D	input:	(None, 199, 32)
	output:	(None, 195, 64)
max_pooling1d_40: MaxPooling1D	input:	(None, 195, 64)
	output:	(None, 39, 64)
conv1d_69: Conv1D	input:	(None, 39, 64)
	output:	(None, 35, 96)
max_pooling1d_41: MaxPooling1D	input:	(None, 35, 96)
	output:	(None, 7, 96)
dropout_51: Dropout	input:	(None, 7, 96)
	output:	(None, 7, 96)
flatten_16: Flatten	input:	(None, 7, 96)
	output:	(None, 672)
dropout_52: Dropout	input:	(None, 672)
	output:	(None, 672)
dense_31: Dense	input:	(None, 672)
	output:	(None, 128)
dropout_53: Dropout	input:	(None, 128)
	output:	(None, 128)
dense_32: Dense	input:	(None, 128)
	output:	(None, 39)

Fig. 2. CNN Model Architecture Summary

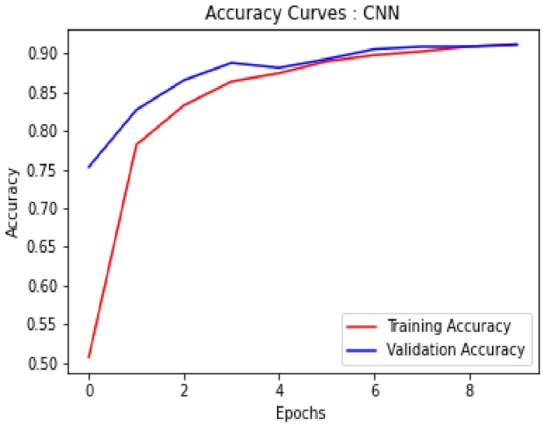


Fig. 3. Accuracy Curve.

The analysis depicts the loss curve observed throughout the model run. Initially, the training phase commenced with a higher loss value of 2.75, gradually diminishing to 0.58 after 10 epochs. Concurrently, the validation phase exhibited consistent performance, with the loss decreasing to 0.427 over the same duration of 10 epochs. These outcomes signify successful learning and convergence of the model during both the training and validation phases.

The proposed method functions akin to traditional CNN classifiers designed for image classification, albeit with the main distinction lying in the input type, which comprises textual data or keywords instead of images. Table 1 provides an overview of the performance of various models in text classifi- cation, with these models sharing a similar implementation but differing primarily in the selection of word embedding. No- tably, the integration of GloVe word vector embedding notably enhances the model's performance. The classification of health records, characterized by specialized medical terminologies, presents inherent challenges. However, the utilization of GloVe word vectors, pre-trained on Wikipedia data, confers distinct advantages owing to Wikipedia's comprehensive coverage of diverse topics, including medical content. This enhancement is evident not only in classifying health record data but also in general text categorization.

Table I: Comparison of Model Accuracy

CNN-GloVe	**CNN-WikiNews**	**RNN-GRU WikiNews**	**BidirectionalRNN WikiNews**
0.87	0.78	0.67	0.68

V. CONCLUSION AND FUTURE WORK

A. Conclusion

In conclusion, this research has demonstrated the efficacy of leveraging advanced deep learning techniques, particularly Convolutional Neural Networks (CNNs) and pre-trained word embeddings like GloVe, for the accurate classification of medi- cal transcription samples. The proposed methodology achieved commendable performance metrics, including high accuracy, precision, recall, and F1 score during validation, underscoring its effectiveness in handling the complexities of health-related text data.

Moreover, the integration of GloVe word vector embed- ding has significantly enhanced the model's performance, particularly in dealing with specialized medical terminologies present in health records. This enhancement not only improves the classification of health record summaries but also holds promise for general text categorization tasks.

B. Future Work

Moving forward, several avenues for future research emerge from this study. Firstly, there is a need to explore the scalability and generalization capabilities of the proposed methodology by applying it to larger and more diverse datasets. This will help assess its robustness across different medical specialties and healthcare contexts.

Additionally, further investigation into the interpretability of the model's predictions is warranted, particularly in healthcare settings where explainability is crucial. Techniques for visu- alizing and understanding the features learned by the model should be explored to provide insights into its decision-making process, thus enhancing trust and usability in clinical practice.

Furthermore, future research should focus on deploying the developed model in real-world healthcare settings. However, this deployment requires careful consideration of data privacy, regulatory compliance, and ongoing model validation and monitoring to ensure its reliability and safety in clinical prac- tice. Collaborations with healthcare institutions and industry partners may facilitate the seamless integration of the model into existing healthcare systems.

REFERENCES

[1] J. Ren, R. Kunes, and F. Doshi-Velez, "Prediction focused topic models for electronic health records," arXiv preprint arXiv:1911.08551, 2019.

[2] M. Hughes, I. Li, S. Kotoulas, and T. Suzumura, "Medical text classifi- cation using convolutional neural networks," in Informatics for Health: Connected Citizen-Led Wellness and Population Health. IOS Press, 2017, pp. 246–250.

[3] S. J. Fodeh, M. Zirkle, D. Finch, R. Reeves, J. Erdos, and C. Brandt, "Medcat: A framework for high level conceptualization of medical notes," in 2013 IEEE 13th International Conference on Data Mining Workshops. IEEE, 2013, pp. 274–280.

[4] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny, "Medex: a medication information extraction system for clinical narratives," Journal of the American Medical Informatics Association, vol. 17, no. 1, pp. 19–24, 2010.

[5] Y. Wu, M. Jiang, J. Xu, D. Zhi, and H. Xu, "Clinical named entity recognition using deep learning models," in AMIA annual symposium proceedings, vol. 2017. American Medical Informatics Association, 2017, p. 1812.

[6] W. Lee, K. Kim, E. Y. Lee, and J. Choi, "Conditional random fields for clinical named entity recognition: a comparative study using korean clinical texts," Computers in biology and medicine, vol. 101, pp. 7–14, 2018.

[7] Z. Zeng, Y. Deng, X. Li, T. Naumann, and Y. Luo, "Natural language processing for ehr-based computational phenotyping," IEEE/ACM trans- actions on computational biology and bioinformatics, vol. 16, no. 1, pp. 139–153, 2018.

[8] R. R. Zavala, P. Martinez et al., "The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: comparative study," JMIR medical informatics, vol. 8, no. 12, p. e18953, 2020.

[9] F. Xie, H. Yuan, Y. Ning, M. E. H. Ong, M. Feng, W. Hsu, B. Chakraborty, and N. Liu, "Deep learning for temporal data represen- tation in electronic health records: A systematic review of challenges and methodologies," Journal of biomedical informatics, vol. 126, p. 103980, 2022.

[10] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun et al., "Scalable and accurate deep learning with electronic health records," NPJ digital medicine, vol. 1, no. 1, pp. 1–10, 2018.

[11] E. Ford, J. A. Carroll, H. E. Smith, D. Scott, and J. A. Cassell, "Extracting information from the text of electronic medical records to improve case detection: a systematic review," Journal of the American Medical Informatics Association, vol. 23, no. 5, pp. 1007–1015, 2016.

[12] Z. Wang, A. D. Shah, A. R. Tate, S. Denaxas, J. Shawe-Taylor,

and H. Hemingway, "Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning," PLoS One, vol. 7, no. 1, p. e30412, 2012.

[13] J. L. Leevy, T. M. Khoshgoftaar, and F. Villanustre, "Survey on rnn and crf models for de-identification of medical free text," Journal of Big Data, vol. 7, no. 1, p. 73, 2020.

[14] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Briefings in bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018.

[15] D. S. Carrell, S. Halgrim, D.-T. Tran, D. S. Buist, J. Chubak, W. W. Chapman, and G. Savova, "Using natural language processing to im- prove efficiency of manual chart abstraction in research: the case of breast cancer recurrence," American journal of epidemiology, vol. 179, no. 6, pp. 749–758, 2014.

[16] S. Kweon, J. H. Lee, Y. Lee, and Y. R. Park, "Personal health infor- mation inference using machine learning on rna expression data from patients with cancer: algorithm validation study," Journal of medical Internet research, vol. 22, no. 8, p. e18387, 2020.

[17] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, and H. Xu, "Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines," Journal of the American Medical Informatics Association, vol. 25, no. 3, pp. 331–336, 2018.

[18] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[19] H.-Y. Kim, J. Lee, N. Y. Yeo, M. Astrid, S.-I. Lee, and Y.-K. Kim, "Cnn based sentence classification with semantic features using word clustering," in 2018 International Conference on Information and Com- munication Technology Convergence (ICTC). IEEE, 2018, pp. 484–488.

[20] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," arXiv preprint arXiv:1510.03820, 2015.