Large language models versus static word embeddings to predict Acute Kidney Injury in the Intensive Care Unit: Does context matter?

Rick van Slobbe[1], Drahomira Herrmannova[2], Elia S. Lima-Walton[2], Ameen Abu-Hanna[3,4] and Iacopo Vagliano[3,4,*]

[1]Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
[2]Elsevier B.V., Amsterdam, the Netherlands
[3]Amsterdam UMC, University of Amsterdam, Amsterdam, the Netherlands
[4]Amsterdam Public Health Research Institute, Amsterdam, the Netherlands

Abstract

We investigated the effect of representing clinical notes using contextualized and static word embeddings. We focus on the early prediction of Acute Kidney Injury (AKI) in the intensive care unit (ICU). We also studied the impact of combining clinical notes with clinical variables. We developed six models based on convolutional neural networks. Our models achieved good predictive performance (AUROC 0.77-0.89). Surprisingly, contextualized and static word embeddings yielded similar performance in all our experiments. Combining text with clinical variables improved the results. Our models can support clinicians in promptly recognizing and treating patients with deteriorating AKI and improving patient outcomes in the ICU.

Keywords

Acute Kidney Injury, Contextualized Word Embeddings, Deep Learning, Natural Language Processing, Clinical Prediction Models

1. Introduction

Increased monitoring of intensive care unit (ICU) patients holds great potential for the early prediction of medical outcomes. Deep learning models have emerged to predict clinical outcomes based on high-dimensional data. In addition to collecting clinical variables' measurements, physicians also write free-text clinical notes during an ICU stay. Clinical notes contain information not recorded via clinical variables, such as previous diagnoses, medication treatments, and disease progression. Traditionally, clinical text has been analyzed using static word embeddings, such as GloVe [1]. More recently, contextualized word embeddings have been developed, also known as large language models (LLMs), such as the Bidirectional Encoder Representations from Transformers (BERT).[1]

Acute kidney injury (AKI) is a sudden reduction in kidney function, measured by increased serum creatinine (SCr) or decreased urine output. AKI is common in the ICU, where it can occur in up to 30% of the stays [2]. Convolutional neural networks (CNNs) proved effective to predict AKI [3, 4, 5].

AKI is a particularly relevant case. SCr and/or urine output defines AKI according to widespread clinical guidelines [6]. However, using these variables as predictors in a prediction model is a form of incorporation bias and can potentially lead to inflated performance. Furthermore, SCr tests are typically required when clinicians already have suspicions of a problem related to the kidneys, while urine output is not commonly measured outside ICU. In the context of early diagnosis, the model can be most useful if it can timely signal the risk of AKI to clinicians anticipating their suspicions. Thus, a better evaluation of AKI models would also assess performance without incorporating SCr and urine output. When such

SLM4Health: Improving healthcare with small language models, 26th June 2025, Pavia, Italy
*Corresponding author.
a.abu-hanna@amsterdamumc.nl (A. Abu-Hanna); i.vagliano@amsterdamumc.nl (I. Vagliano)
0000-0002-2730-1546 (D. Herrmannova); 0000-0003-4324-7954 (A. Abu-Hanna); 0000-0002-3066-9464 (I. Vagliano)

[1]In the natural language processing community there is currently no agreement on an LLM definition. We consider large language models from the transformers onward.

variables are not included, the models would most profit from relying on clinical notes rather than structured clinical variables, and/or from combining clinical notes with clinical variables. To the best of our knowledge, an in-depth analysis of AKI models under these conditions has not yet been conducted.

To address all these problems and conduct such an analysis, we focus on predicting whether a patient has AKI or will develop AKI during the ICU stay. The prediction occurs after the first 48 hours of a stay. More specifically, we address the following research question: *What effect do contextualized word embeddings have on the predictive performance compared to static word embeddings to predict AKI within the ICU stay based on the first 48 hours of the stay?*

To answer this question, we develop six CNN models to process individual and multiple data modalities (unstructured clinical notes and structured clinical variables) for predicting AKI within the ICU stay. We compare static and contextualized embeddings using GloVe and BERT, respectively. In a previous work, we investigated multi-modal models to predict AKI [7]. This study provides the first in-depth comparison between models using different word representations. Despite LLMs' potential, smaller embedding might be a promising alternative in clinical settings due to their adaptability, efficiency, and lower resource demands, especially in resource-constrained environments and/or when the performance gap is small. We also both include and exclude SCr and urine output.

2. Related work

Various work on AKI prediction in the ICU using machine learning has been published [8]. Most studies used only clinical variables as input for the proposed models. Sato et al. proposed a CNN architecture based on one-dimensional convolutional filters [3] using structured clinical data from the first 48 hours of the ICU stay. Vagliano et al. developed and externally validated interpretable and continuous models for AKI in the ICU [9, 10].

Other studies have leveraged unstructured clinical notes. Le et al. [4] developed a CNN that used data from the first 53 hours of the ICU stay to predict AKI up to 48 hours before onset. Li et al. proposed multiple models to predict AKI based on clinical notes from the first 24 hours of the ICU stay [5].

Some studies also incorporated domain knowledge into their text representations. Vagliano et al. [11] compared using only clinical variables, versus clinical variables and notes for predicting AKI. Notes were also enhanced with external knowledge from the Unified Medical Language System (UMLS) and SNOMED CT knowledge graphs. Brancato et al.[12] further extended the model proposed in [11] to handle clinical notes enriched with multi-word concepts (in contrast to the single-word concepts used before).

The studies previously listed relied on different tasks (e.g., predicting AKI after the first 24 vs after the first 48 hours), preprocessing, and internal validation strategies (e.g., splitting into training validation and test sets versus cross-validation versus bootstrapping, different splits, folds, etc.). Given this method variation, their performances are not directly comparable even when using the same data. In contrast, we provide directly comparable results for various modalities. We do so both when including and excluding SCr and urine output, while previous studies included them. To the best of our knowledge, such an in-depth comparison for the prediction of AKI has not been performed.

Recently, among LLMs, transformer architectures have become increasingly popular for many general text classification tasks with substantial performance increases compared to general machine learning methods. Mao et al. and Li et al. investigated the efficacy of using BERT in the disease-specific domain of early AKI prediction [13, 14]. To the best of our knowledge, (1) there is currently no research that compared static and contextualized embeddings of clinical notes for the prediction of AKI and (2) no previous work has proposed a multi-modal model based on BERT to predict AKI.

3. Methods

3.1. Data, patient cohort and outcome definition

The data used to develop the model was from the MIMIC-III database [15], which holds de-identified data of ICU stays in a US hospital between 2001 and 2012. Data includes vital sign measurements, laboratory results, and clinical notes. We included patients older than 18 years whose stays contained at least one measurement of SCr or urine output. After applying the inclusion criteria, 44,303 stays of 32,664 unique patients were selected. The patients' characteristics can be found in Table 1. The outcome was whether patients developed AKI during their stay in the ICU. AKI was defined via the KDIGO guidelines [6].

Table 1
Characteristics of MIMIC population [7]. Continuous variables are given as median and interquartile range. Categorical variables are displayed as absolute counts and percentages of stays.

Characteristics	Overall	No AKI	AKI	Missing (%)
Demographics and administrative variables				
Number of patients	32,664 (100.00%)	11,439 (35.02%)	23,906 (64.98%)	
Number of stays	44,303 (100.00%)	12,971 (29.28%)	31,332 (70.72%)	
Males	25,210 (56.90%)	7,322 (56.44%)	17,888 (57.98%)	0.00
Age (years)	66.09 [53.44-77.73]	61.12 [47.80-74.87]	67.85 [55.82-78.55]	0.00
Admission type				0.00
Emergency	36,613 (82.64%)	11,100 (85.58%)	25,513 (81.43%)	
Elective	6,495 (14.66%)	1,568 (12.09%)	4,927 (15.73%)	
Urgent	1,195 (2.70%)	303 (2.34%)	892 (2.85%)	
Ethnicity				0.00
White	31,948 (72.11%)	9,360 (72.16%)	22,588 (72.09%)	
Unknown	4,735 (10.69%)	1,183 (9.12%)	3,552 (11.34%)	
Black	3,973 (8.97%)	1,103 (8.50%)	2,870 (9.16%)	
Hispanic	1,461 (3.30%)	541 (4.17%)	920 (2.94%)	
Other	1,121 (2.53%)	421 (3.25%)	756 (2.41%)	
Asian	1,040 (2.35%)	365 (2.81%)	628 (2.00%)	
Native	25 (0.06%)	7 (0.05%)	18 (0.06%)	
Vitals and laboratory results				
Anion gap (mEq/L)	13.00 [12.00 - 16.00]	13.00 [11.00-15.00]	14.00 [12.00-16.00]	93.75
Bicarbonate (mg/dL)	25.00 [22.00 - 28.00]	25.00 [23.00-28.00]	25.00 [22.00-27.00]	93.49
Blood urea nitrogen (mg/dL)	20.00 [13.00 - 33.00]	15.00 [11.00-23.00]	22.00 [15.00-37.00]	93.41
Chloride (mEq/L)	106.00 [102.00 - 110.00]	106.00 [103.00-109.00]	106.00 [102.00-110.00]	92.84
Diastolic blood pressure (mmHg)	58.50 [50.60 - 68.00]	61.00 [52.00-70.00]	58.00 [50.00-67.00]	21.88
Glucose (mg/dL)	126.00 [105.00 - 158.00]	123.00 [103.00-153.00]	128.00 [106.00-160.00]	77.49
Hematocrit (%)	29.80 [26.50 - 33.50]	30.70 [27.20-34.60]	29.40 [26.20-33.00]	90.61
Hemoglobin (g/dL)	10.70 [9.60 - 12.00]	11.00 [9.80-12.40]	10.60 [9.50-11.80]	92.80
Heart rate (beats/m)	84.00 [72.50 - 96.00]	82.00 [71.00-95.00]	84.00 [73.00-96.00]	19.89
SCr (mg/dL)	1.00 [0.70 - 1.50]	0.80 [0.60-1.00]	1.00 [0.80-1.70]	93.37
Sodium (mEq/L)	139.00 [137.00 - 142.00]	139.50 [137.00-142.00]	139.00 [136.00-142.00]	92.40
$SpO2$[2] (%)	98.00 [96.00 - 99.50]	98.00 [96.00-99.00]	98.00 [96.00-99.75]	22.96
Systolic blood pressure (mmHg)	117.00 [104.00 - 133.00]	119.00 [106.00-134.00]	116.00 [103.00-132.00]	21.87
Potassium (mEq/L)	4.00 [3.60 - 4.30]	3.90 [3.60-4.20]	4.00 [3.70-4.40]	90.07
Respiratory rate (breaths/m)	18.00 [15.00 - 22.00]	18.00 [15.00-21.00]	18.00 [15.00-22.00]	21.38
Urine output 6hr (mL/kg/h)	0.86 [0.42 - 1.69]	1.37 [0.74-2.46]	0.71 [0.38-1.37]	47.62
Urine output 12hr (mL/kg/h)	0.87 [0.44 - 1.58]	1.28 [0.74-2.10]	0.73 [0.40-1.33]	47.62
Urine output 24hr (mL/kg/h)	0.95 [0.49 - 1.66]	1.36 [0.79-2.18]	0.82 [0.45-1.42]	47.62
White blood cell count (K/μL)	11.50 [8.30 - 15.50]	10.60 [7.70-14.30]	11.80 [8.60-16.00]	93.90
Treatments and medications				
Mechanical ventilation	9,216 (20.80%)	1,764 (13.60%)	7,452 (31.17%)	0.00
Sedatives	5,095 (11.50%)	1,005 (7.75%)	4,090 (17.11%)	0.00
Vasopressors	3,543 (8.00%)	595 (4.59%)	2,948 (12.33%)	0.00

[2]Peripheral capillary oxygen saturation

3.2. Text preprocessing and word embeddings

Clinical notes include clinician progress notes, nursing assessments, radiology reports, and laboratory results. Discharge summaries were discarded to prevent information leakage from outside the first 48 hours of the stay. Only text from the first 48 hours of the stay was used. Table 2 outlines the descriptive statistics of the selected clinical notes.

Table 2
Descriptive statistics of clinical notes [7].

Description	Value
Total words	38,028,009
Total unique words	346,783
Total sentences	4,577,993
Average words per admission	858.36 $\pm$ 415.94
Average unique words per admission	396.85 $\pm$ 162.92
Average sentences per admission	103.33 $\pm$ 47.46
Average words per sentence	8.61 $\pm$ 1.91

GloVe embeddings were trained on all selected clinical notes. To ensure a fixed length input, shorter clinical notes were padded and longer ones were truncated keeping only the first 24 words of each of the first 150 sentences. About 90% of the sentences contained less than 24 words, and roughly the great majority of the notes contained less than 150 sentences. Using more sentences and words did not improve performance. The resulting vectors represent words in a 100-dimensional vector space, capturing their semantic relationships. Only words with more than 5 occurrences were kept to limit unreliable representations of rare words.

BERT models require tokenization of words before being input to the model, this process involves breaking down a sequence of words into smaller units called tokens, which can be words, subwords, or characters. For our BERT model, we used 1,536 tokens at most. We considered multiples of 512 as BERT imposes a maximum input sequence length of 512 tokens. We split the 1,536-tokens-long clinical notes into three 512-token chunks. About 85% of notes contained less than 1,536 words. Using more tokens did not improve performance. To ensure a fixed-length input, shorter clinical notes were padded and longer notes were truncated keeping the first 1,536 tokens.

3.3. Preprocessing of structured data

We selected the same 35 clinical variables and applied the same preprocessing as in our previous work [7]. These variables covered a variety of factors, including temporal measurements such as heart rate and platelet count, as well as categorical information such as sex, ethnicity, and the urgency of stay in the ICU. The full list of variables can be found in another study that used the same population, variables and preprocessing [7].

The categorical variables were encoded using the one-hot encoding approach, forming a separate binary feature for each value of the original categorical feature. Temporal features were resampled to 1-hour intervals with a mean aggregation. Variables with less than 50% missing values were imputed by carrying forward the last available value. Variables with 50% missing values or more were imputed using mean imputation. Data was capped at the 1^{st} and 99^{th} percentiles to filter out outliers, and normalized to the [0, 1] range using min-max scaling: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$, where x is the original value and x' is the normalized value. Only selected predictors from the first 48 hours of each stay were used, independent of whether AKI occurred afterward.

3.4. Model development

3.4.1. Models with clinical notes

To predict AKI using clinical notes, a CNN architecture (GloVe-CNN) was developed (Figure 1) [7]. GloVe-CNN incorporated a GloVe embedding layer to represent words as 100-dimensional vectors. Two convolutional layers for feature extraction followed this embedding layer. The first layer independently maps sentences to sentence embeddings, while the second layer combines sentences into a single patient representation. For both levels, we used max pooling. A dropout layer for regularization follows the two convolutional layers, and then a classification layer to predict the probability that a patient has or will develop AKI. GloVe-CNN adopted target replication, which allows the computation of the loss at the sentence level [16]. Through it, GloVe-CNN effectively learned sentence representations tailored for the AKI prediction.

Clinical BioBERT was fine-tuned to predict AKI [17]. The Clinical BioBERT model followed the standard BERT architecture [18] and included an additional classification layer. Max pooling was applied to the 768-dimensional outputs to retain the most important features. A final 768-dimensional vector was obtained and used to fine-tune the Clinical BioBERT model for our task. We also tried using BERT embeddings with the same CNN architecture used for GloVe but it performed worse than the standard BERT architecture.

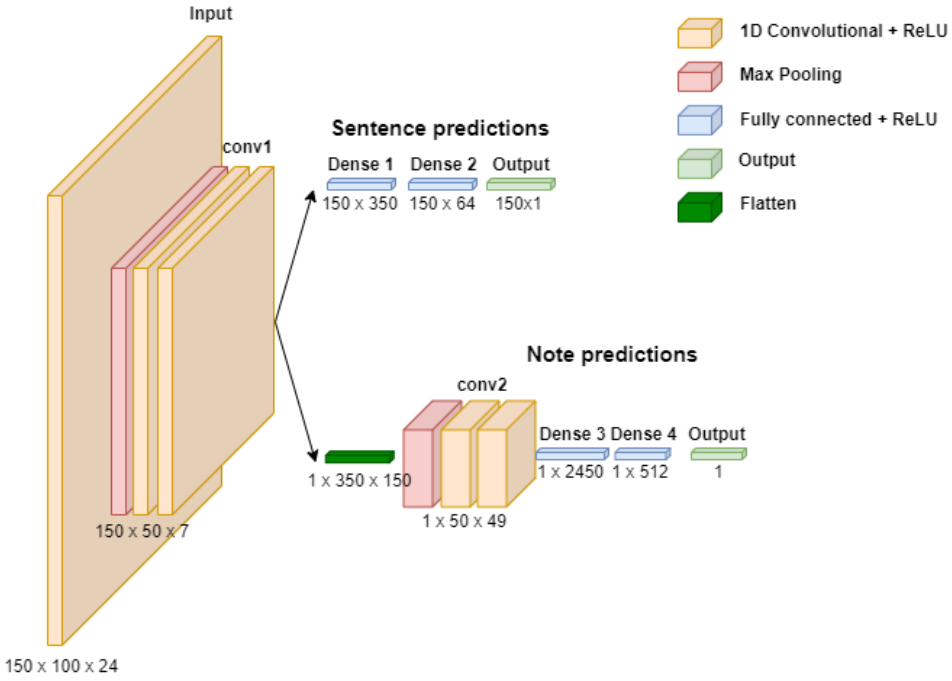

Figure 1: The GloVe-CNN architecture [7]. The convolutional layers are one-dimensional filters with kernel sizes 3, padding with ones, and stride of one.

3.4.2. Extraction of clinical variables

To extract clinical variables and generate a (partial) patient representation from them, a CNN (Var-CNN) was developed (Figure 2) [7]. Var-CNN employed two-dimensional filters and incorporated three stacked convolutional layers, with a subsequent 3x3 max-pooling operation following the final convolutional layer. We relied on two-dimensional filters instead of the more typical one-dimensional filters used with tabular data. This decision was based on initial tests that revealed that two-dimensional filters consistently outperformed one-dimensional filters for this task [7].

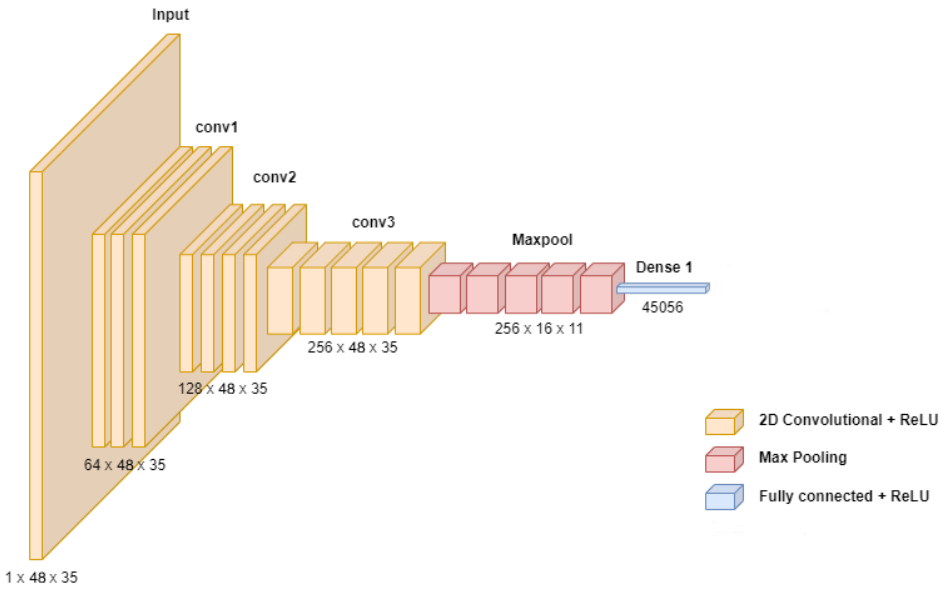

Figure 2: The Var-CNN architecture [7]. The input dimensions correspond to 35 variables and 48 time steps. The convolutional layers are two-dimensional filters with kernel sizes 3x3, padding with ones, and stride of one. The number of filters is doubled in each subsequent convolutional layer.

3.4.3. Multi-modal model

The multi-modal network (Figure 3) builds upon the single-modality models with an intermediate fusion approach. We combined modalities after CNN feature extraction and the first linear layer, or at the last layer of the BERT architecture, excluding the additional classification layer. Each modality branch contributed an output representation, which was then concatenated with the others and passed through a fully connected layer to obtain the final prediction.

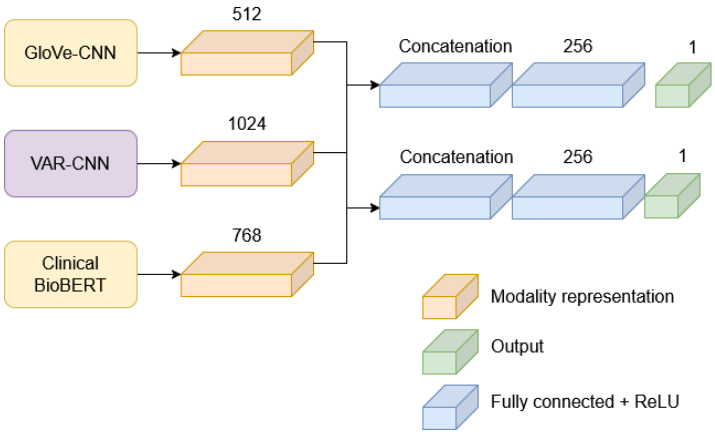

Figure 3: Multi-modal models' architecture. The modality representations are extracted from the second linear layer after convolutions of the CNNs. The VAR-CNN is used individually with GloVe-CNN and Clinical BioBERT, respectively.

3.5. Internal validation

We split the data into 80%, 10%, and 10% for the training, validation and test set, respectively. Both the GloVe-CNN and Clinical BioBERT were evaluated when using only clinical notes and when using clinical notes in combination with clinical variables to assess the performance in a single- vs multi-modal setting. The predictive performance was measured by the area under the receiver operating curve

(AUROC), the area under the precision-recall curve (AUPRC), and the Brier score.

The hyperparameters of the models were tuned on the validation set via grid search. The GloVe-CNN model was set to have 50 filters, as in [16]. For all models, we tested max pooling kernels of size 2, 3, and 4 as well as learning rates of 0.01, 0.001, and 0.0001. The experiments with clinical variables were performed with the number of initial filters of the Var-CNN varying from 8 to 64 with an incremental step of 8. The learning rate was set to 0.0001 for the Var-CNN and all multi-modal models, and to 0.001 for the GloVe-CNN.

4. Results

The performance metrics for predictive models using BERT and GloVe embedding methods are displayed in Table 3. The findings are shown for models that exclusively utilize clinical notes, as well as models that incorporate clinical notes in conjunction with clinical variables, with and without the inclusion of SCr and urine output. The results in Table 3 indicate that similar performance metrics are obtained for models using contextualized and static word embeddings for all combinations of modalities. Glove was selected for further experiments based on these results because training those embeddings requires a 20 times lower computational time than training BERT embeddings.

Table 3
Performance on the MIMIC population for GloVe and BERT models for single- and multi-modality configurations.

Model	AUROC	AUPRC	Brier score
BERT	0.7692	0.8694	0.1771
GloVe	0.7747	0.8881	0.1705
With SCr and urine output			
VAR + BERT	0.8868	0.9525	0.1240
VAR + GloVe	0.8893	0.9531	0.1248
Without SCr and urine output			
VAR + BERT	0.8071	0.9082	0.1578
VAR + GloVe	0.8058	0.9058	0.1592

5. Discussion

This study concerned the effects on the predictive performance of using contextualized word embeddings (LLMs) compared to static word embeddings. Our results showed no differences between the GloVe and BERT approaches (Table 3). Gao et al. came to a similar conclusion that models such as word-level CNNs and hierarchical self-attention networks often perform equally or better compared to BERT models for medical classification tasks [19]. They argue that these results can be attributed to two main reasons. First, in the context of clinical text classification, a small subset of words is highly important for the specific task, while the remaining text can be considered as noise. However, BERT models inherently struggle to focus on specific biomedical keywords due to their nature to instead focus on complex word relationships [19]. This focus on word relationships shifts BERT's attention away from these keywords [19]. The second reason is that the tokenizer employed for BERT models splits unknown words into subwords. While static word embeddings can learn specific keywords important for the given task, BERT is unable to do so because keywords are often split into subwords. BERT models need to discern accurate labels for each subword rather than single keywords, increasing the complexity of the task.

Previous studies have also shown that as the amount of training data increases, the performance gap between static and contextualized word embeddings shrinks [20]. This might indicate that the amount

of training data used in this study was sufficient for the static word embeddings to perform equally well compared to the contextualized word embeddings.

The main strenght is the internal validation of all modality combinations without the SCr and urine output variables. These variables were used to define the AKI label, which introduced a form of incorporation bias. Evaluation without these variables offers information on the comparison of different modalities without this incorporation bias. There are also limitations. First, internal validation was based on a simple split into training, validation, and test sets. While cross-validation or bootstrap are better validation methods, using these methods to train Clinical BioBERT was unfeasible in our environment. Second, GloVe embeddings and the Clinical BioBERT model are pre-trained on the MIMIC-III dataset. The issue of data leakage is mitigated since no labels were exposed during the pre-training phase. Since GloVe and the clinical BioBERT had encountered the test data before, the model's performance might be overestimated. We trained GloVe on all the clinical notes to fairly compare it with Clinical BioBERT.

6. Conclusions and future work

We compared using LLMs versus static word embeddings to predict AKI in the ICU when using only clinical notes and when combining clinical variables and notes. We also showed the effect of including versus excluding SCr and urine output. Surprisingly, static word embeddings performed similarly to LLMs, both when including and excluding clinical variables as well as when including and excluding SCr and urine output. Combining text with clinical variables improved the results.

Our models can support clinicians to promptly recognize and treat patients with deteriorating AKI and consequently improve patient outcomes in the ICU. Our extensive comparison of modalities and text representations may further guide researchers and practitioners in leveraging multi-modal models to predict AKI and inspire them to investigate multi-modality and LLMs for other tasks.

Acknowledgments

The computational resources used were financed by the NWO programme Computing Time on National Computer Facilities (grant 2024.15).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation., in: EMNLP, volume 14, 2014, pp. 1532–1543.

[2] J. Case, S. Khan, R. Khalid, A. Khan, et al., Epidemiology of acute kidney injury in the intensive care unit, Critical care research and practice 2013 (2013).

[3] N. Sato, E. Uchino, R. Kojima, S. Hiragi, M. Yanagita, Y. Okuno, Prediction and visualization of acute kidney injury in intensive care unit using one-dimensional convolutional neural networks based on routinely collected data, Computer Methods and Programs in Biomedicine 206 (2021) 106129.

[4] S. Le, A. Allen, J. Calvert, P. M. Palevsky, G. Braden, S. Patel, E. Pellegrini, A. Green-Saxena, J. Hoffman, R. Das, Convolutional neural network model for intensive care unit acute kidney injury prediction, Kid Int Rep 6 (2021).

[5] Y. Li, L. Yao, C. Mao, A. Srivastava, X. Jiang, Y. Luo, Early prediction of acute kidney injury in critical care setting using clinical notes, in: BIBM, 2018.

[6] A. Khwaja, Kdigo clinical practice guidelines for acute kidney injury, Nephron Clinical Practice 120 (2012) c179–c184.
[7] R. van Slobbe, D. Herrmannova, D. Boeke, E. Lima-Walton, A. Abu-Hanna, I. Vagliano, Multimodal convolutional neural networks for the prediction of acute kidney injury in the intensive care, International Journal of Medical Informatics 196 (2025) 105815.
[8] I. Vagliano, N. C. Chesnaye, J. H. Leopold, K. J. Jager, A. Abu-Hanna, M. C. Schut, Machine learning models for predicting acute kidney injury: a systematic review and critical appraisal, Clinical Kidney Journal 15 (2022) 2266–2280.
[9] I. Vagliano, O. Lvova, M. C. Schut, Interpretable and continuous prediction of acute kidney injury in the intensive care, in: Public Health and Informatics, volume 281, 2021, pp. 103–107.
[10] I. Vagliano, C. Byrne Salsas, T. Wünn, M. C. Schut, External validation and transportability of models to predict acute kidney injury in the intensive care unit, in: Inf and Tech in Clin Care and Public Health, volume 295, 2022, pp. 148–151.
[11] I. Vagliano, W.-H. Hsu, M. C. Schut, Machine learning, clinical notes and knowledge graphs for early prediction of acute kidney injury in the intensive care, Stud. Health Technol. Inform 289 (2022) 329–332.
[12] L. Brancato, I. Calixto, A. Abu-Hanna, I. Vagliano, Leveraging multi-word concepts to predict acute kidney injury in intensive care, in: Healthcare Transformation with Informatics and Artificial Intelligence, volume 305, 2023, pp. 10–13.
[13] C. Mao, L. Yao, Y. Luo, Aki-bert: a pre-trained clinical language model for early prediction of acute kidney injury, arXiv:2205.03695 (2022).
[14] Y. Li, R. M. Wehbe, F. S. Ahmad, H. Wang, Y. Luo, Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences, arXiv:2201.11838 (2022).
[15] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific data 3 (2016) 1–9.
[16] P. Grnarova, F. Schmidt, S. L. Hyland, C. Eickhoff, Neural document embeddings for intensive care patient mortality prediction, arXiv:1612.00467 (2016).
[17] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, arXiv:1904.03323 (2019).
[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.
[19] S. Gao, M. Alawad, M. T. Young, J. Gounley, N. Schaefferkoetter, H. J. Yoon, X.-C. Wu, E. B. Durbin, J. Doherty, A. Stroup, et al., Limitations of transformers on clinical text classification, J Biomed and Health Inform 25 (2021) 3596–3607.
[20] S. Arora, A. May, J. Zhang, C. Ré, Contextual embeddings: When are they worth it?, arXiv:2005.09117 (2020).