VER
Vascular & Endovascular Review

Natural Language Processing (NLP) for Diabetes Research: Identifying Hidden Risk Patterns in Electronic Health Records

Dr. S. Revathi[1], Dr. K. Arpitha[2], Dr. Ujwalla Gawande[3], Dr. Ashish Kumar Tamrakar[4], N Gold Pearlin Mary[5], Deepa Abin[6]

[1]Designation: Professor Institute: B.S.Abdur Rahman Crescent Institute of Science and Technology District: Chingleput City : Chennai State: Tamil Nadu
Email: srevathi@crescent.education
[2]Designation: Associate professor Department:CSE(AIML Institute: Geetanjali college of engineering and technology District: cheeryal
City:keesara, State:Telangana
Email -drarpita.cse@gcet.edu.in
[3]Professor and Dean R & D, Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra. ujwallgawande@yahoo.co.in
[4]Designation: Associate Professor Department: Computer Science & Engineering Institute:RSR, Rungta College of Engineering & Technology
District:Durg City: Durg State: Chhattisgarh
Mail id: ashish.tamrakar1987@gmail.com
[5]Professor, College address with pin code: Sree Balaji Dental College and Hospital , Pallikaranai, chennai - 600100
drngoldpearlinmary@gmail.com
9962105023 Orcid id – 0000-0002-2728-1493
[6]Associate Professor Department of CSE-Data Science Vishwakarma Institute of Technology Pune Maharashtra
deepaabin@gmail.com

ABSTRACT

Diabetes continues to rise globally, driven by complex interactions of genetics, lifestyle, and healthcare disparities. Traditional statistical approaches often fail to capture subtle clinical clues buried in “Electronic Health Records (EHRs)”, including physician notes, diagnostic histories, medication patterns, and temporal disease trajectories. This study proposes an integrated “Natural Language Processing (NLP)” framework for identifying hidden risk patterns associated with early diabetes progression and related complications. The system processes unstructured clinical text from EHRs using transformer-based embeddings, temporal sequence modelling, and clinical concept extraction to reveal latent predictors of glycemic deterioration. By combining contextual word representations with risk-stratification models, the framework uncovers high-impact patterns such as symptom clustering, comorbidity interactions, medication response inconsistencies, and consultation frequencies. Experiments conducted on anonymised hospital datasets show substantial improvements in early-risk prediction accuracy, interpretability, and longitudinal monitoring compared to traditional machine-learning baselines. Moreover, the proposed model demonstrates efficiency in extracting clinically relevant biomarkers without manual feature engineering. The study highlights the potential of NLP-driven analytics for supporting personalised interventions, enhancing clinical decision support systems, and accelerating precision-diabetes research. The results provide evidence that leveraging unstructured EHR text through advanced NLP architectures is a transformative pathway for early detection, complication forecasting, and population-level diabetes surveillance.

KEYWORDS: Diabetes, Electronic Health Records, Natural Language Processing, Risk Prediction, Clinical Text Mining, Transformer Models.

How to Cite: S. Revathi1, K. Arpitha, Ujwalla Gawande, Ashish Kumar Tamrakar, N Gold Pearlin Mary, Deepa Abin., (2025) Natural Language Processing (NLP) for Diabetes Research: Identifying Hidden Risk Patterns in Electronic Health Records, Vascular and Endovascular Review, Vol.8, No.15s, 379-384

INTRODUCTION

The rapid growth of diabetes worldwide has created a pressing need for early identification of high-risk individuals. Healthcare organisations are overwhelmed with electronic health data generated across outpatient visits, laboratory results, diagnostics, prescriptions, and follow-up consultations. A considerable proportion of this information exists in unstructured text, including physician narratives, progress notes, discharge summaries, and radiology interpretations. These textual components contain rich contextual insights into patient symptoms, behavioural indicators, medication adherence, and long-term disease progression. Yet, traditional analytical models rely primarily on structured fields such as lab values and coded diagnoses, overlooking massive reservoirs of clinically meaningful knowledge embedded in narrative data. “Natural Language Processing (NLP)” provides an opportunity to systematically analyse this unstructured data, extract hidden risk signals, and identify patterns unavailable in conventional datasets.

Recent advancements in deep learning, particularly transformer-based architectures, have enabled NLP systems to capture semantic nuances, temporal dependencies, and clinical correlations with unprecedented accuracy. These models can analyse complex linguistic cues, identify early indicators of metabolic dysfunction, and detect risk trajectories long before traditional screening tools. For diabetes specifically, early detection of complications such as neuropathy, nephropathy, cardiovascular deterioration, and foot ulcers relies heavily on subtle textual patterns often found in physician notes. Existing diabetes prediction models focus mainly on numeric indicators like HbA1c levels, BMI, or glucose readings. However, clinical narratives may reference lifestyle changes, recurrent symptoms, familial predispositions, dietary challenges, or medication compliance issues that are not represented in structured fields. By integrating unstructured and structured EHR data, NLP-based models offer a powerful solution for uncovering latent predictors of diabetes onset and progression.

This research proposes a comprehensive NLP-driven framework designed to identify hidden risk patterns in diabetes-related EHRs. The study introduces clinical concept extraction pipelines, context-aware embeddings, and temporal modelling to dissect the linguistic characteristics of patient histories. The approach emphasises semantic clustering of symptoms, predictive modelling of complication risks, and dynamic analysis of textual trends over time. By validating the framework on large-scale anonymised datasets, the study demonstrates significant improvements in predictive accuracy, interpretability, and adaptability across diverse clinical settings. Ultimately, this work contributes toward building next-generation intelligent systems capable of supporting clinicians in proactive diabetes management and strengthening the foundation for precision-medicine applications in chronic disease research.

RELATED WORKS

The application of NLP to healthcare has expanded rapidly, especially in disease prediction, clinical text mining, and automated phenotyping. Earlier studies primarily relied on keyword extraction and rule-based systems, which provided limited contextual understanding and struggled with linguistic ambiguity [1]. Classical machine-learning models applied techniques such as bag-of-words and TF-IDF representations to clinical text, enabling basic classification tasks but falling short in capturing complex clinical semantics [2]. As diabetes research evolved, various studies attempted to correlate textual features from EHRs with glycemic outcomes, but the predictive capabilities remained constrained by shallow feature representations [3].

The introduction of distributed word embeddings, such as Word2Vec and GloVe, significantly improved representation quality by encoding semantic relationships within clinical narratives [4]. Subsequent research advanced to domain-specific embeddings like BioWordVec and ClinicalBERT, which demonstrated improved performance in extracting medical terminologies and contextual cues relevant to chronic diseases [5]. Several works explored diabetes detection using NLP-driven phenotyping, where clinical concepts extracted from provider notes were combined with structured lab data [6]. For example, Ling et al. developed text-mining pipelines that identified early neuropathy indicators through physician notes, achieving greater sensitivity than diagnostic code-based models [7]. Similarly, NLP-enabled clinical decision support systems have been tested to predict retinopathy progression by analysing ophthalmology reports, showcasing strong potential in identifying early-stage risks [8].

Recent advances in transformer models have shifted the landscape of medical NLP. Studies employing ClinicalBERT, BlueBERT, and BioGPT have demonstrated superior performance in capturing nuanced risk factors and temporal linguistic patterns from EHR narratives [9]. Furthermore, integrated multimodal architectures combining structured and unstructured data have been explored for chronic disease management [10]. In diabetes-specific literature, transformer-based approaches have enabled extraction of subtle behavioural patterns such as diet adherence references, exercise irregularities, and emotional cues associated with diabetes management challenges [11]. These findings indicate that textual features can serve as powerful predictors of long-term metabolic deterioration.

In addition to disease prediction, NLP has been applied to comorbidity detection, complication forecasting, and patient stratification. Studies integrating NLP with probabilistic models have shown improved capabilities for identifying undiagnosed diabetes by analysing symptom clusters within narrative notes [12]. Temporal NLP models have also been developed to track disease progression using longitudinal health records, highlighting the importance of capturing sequential patterns in clinical contexts [13]. Despite these advancements, gaps remain in integrating semantic extraction, temporal modelling, and risk-stratification into unified frameworks tailored specifically for diabetes. Existing works rarely explore dynamic associations among symptoms, comorbidities, medications, and lifestyle factors captured within clinical narratives [14]. The present research addresses these gaps by proposing a unified NLP pipeline that identifies hidden, multi-dimensional risk patterns across diverse EHR texts, advancing the predictive capabilities for diabetes-related outcomes [15].

METHODOLOGY

3.1 Research Design

This study employs an integrated methodological approach that combines clinical text mining, deep learning-based NLP models, and risk-stratification analytics. The overall goal is to identify hidden diabetes risk patterns embedded within unstructured EHR narratives. The research design incorporates advanced transformer embeddings, sequential modelling, and semantic clustering to extract both explicit and implicit indicators of diabetes progression. The pipeline emphasises robust preprocessing, contextual feature extraction, and predictive modelling to evaluate the effectiveness of NLP-driven risk identification across real-world EHR datasets. The design further prioritises interpretability through attention-based mechanisms and concept-level mapping to unify linguistic signals with clinical relevance.

3.2 Study Data Source

The study utilises anonymised EHR data obtained from multi-specialty hospitals, consisting of physician notes, discharge summaries, outpatient consultation entries, medication histories, and laboratory reports. The dataset includes patients diagnosed with diabetes, prediabetes, and non-diabetic controls. Key components extracted include symptom mentions, comorbidity interactions, lifestyle descriptions, longitudinal follow-up narratives, and medication-related patterns. This diversity of unstructured text facilitates detailed exploration of hidden risk dynamics associated with diabetes progression.

Table 1: Dataset Composition

Component Type	Quantity	Description
Physician Notes	45,000	Narrative descriptions of symptoms and assessments
Discharge Summaries	18,500	Hospitalisation outcomes and recommendations
OPD Consultation Notes	33,200	Follow-up observations and lifestyle discussions
Medication Histories	27,400	Drug adherence, dosage adjustments, side effects
Lab Report Summaries	22,800	Textual interpretations of diagnostic values

3.3 Data Preprocessing

Preprocessing involved de-identification, sentence segmentation, negation detection, medical concept extraction, and token normalisation. Clinical concepts such as symptoms, comorbidities, medications, and lifestyle behaviours were mapped to UMLS and SNOMED-CT vocabularies. Bi-gram and tri-gram medical phrase merging was applied to preserve multi-word clinical entities (e.g., “foot ulcer,” “peripheral neuropathy”). Noise reduction steps included removal of non-clinical entries, abbreviation expansion, and contextual normalisation to ensure linguistic consistency for downstream modelling.

3.4 NLP Feature Extraction

The framework employed transformer-based embeddings derived from ClinicalBERT and BioGPT to capture contextual semantics. Attention-based concept extraction was used to highlight high-impact linguistic markers, while temporal embedding modules captured risk progression across longitudinal patient timelines. These embeddings were concatenated with structured lab indicators such as HbA1c trajectories to enhance multimodal risk modelling.

Table 2: NLP Feature Extraction Parameters

Feature Category	Technique Applied	Description
Context Embedding	ClinicalBERT, BioGPT	Captures semantic and contextual meaning
Medical Concept Extraction	UMLS/SNOMED Mapping	Identifies clinical entities and relationships
Temporal Encoding	Positional + Sequential Layer	Captures progression and longitudinal patterns
Attention Weights	Multi-head Attention	Highlights high-impact textual regions

3.5 Risk Prediction Model Architecture

The proposed risk prediction framework employs a hybrid deep-learning architecture designed to capture both semantic richness and temporal progression within clinical narratives. At its core, the system integrates transformer-based contextual embeddings with recurrent sequence modelling to build a multi-dimensional understanding of patient risk indicators. Transformer embeddings derived from ClinicalBERT and BioGPT form the foundation of the model, enabling the extraction of nuanced linguistic representations that reflect symptom descriptions, clinician interpretations, behavioural observations, and implicit cues often overlooked in structured EHR fields. These embeddings effectively encode long-range dependencies and domain-specific semantics essential for interpreting complex medical text. To complement this contextual understanding, Bidirectional "Long Short-Term Memory (BiLSTM)" layers are incorporated to model the sequential and temporal relationships across narrative entries. This component captures progression patterns embedded within longitudinal records, such as repeated symptom mentions, evolving medication responses, and shifting lifestyle behaviours. The bidirectional design enables the model to consider both prior and subsequent clinical events, offering a more coherent representation of diabetes risk evolution over time. The output from the BiLSTM layers is fused with structured clinical indicators, including lab-derived features like HbA1c trends, blood pressure histories, and medication dosages when available. This multimodal fusion creates a unified feature space where textual and numerical elements interact, improving the capacity to detect subtle risk signatures. A series of fully connected dense layers further refine these representations, applying nonlinear transformations to separate high-risk and low-risk profiles while filtering out noise in the feature space. The final risk-classification head produces calibrated probability scores indicating the likelihood of early diabetes onset, complication development, or rapid metabolic decline. These scores are generated through a sigmoid or softmax layer depending on the prediction task. Attention mechanisms embedded within the architecture highlight influential tokens, enabling the system to prioritise clinically significant concepts such as symptom clusters, comorbidity interactions, lifestyle deviations, and medication adherence discrepancies. This layered design ensures that risk predictions are not only accurate but also interpretable and clinically meaningful, providing a strong foundation for proactive diabetes management and personalised care pathways.

3.6 Evaluation Metrics

Model performance was evaluated using a comprehensive suite of metrics designed to capture both predictive accuracy and clinical reliability. The primary measures included F1-score, precision, recall, and AUC-ROC, each offering distinct insight into the model's ability to correctly identify high-risk diabetes cases while minimising false alarms. Precision quantified the correctness of positive predictions, ensuring that flagged risk indicators corresponded to actual clinical concerns, whereas recall assessed the model's sensitivity in capturing subtle linguistic cues signalling early metabolic deterioration. The F1-score, as the harmonic mean of precision and recall, provided a balanced measure particularly suitable for datasets with class imbalance, which is common in chronic-disease prediction.

AUC-ROC was used to evaluate the trade-off between true-positive and false-positive rates across varying decision thresholds, offering a robust view of how reliably the model distinguished at-risk individuals from controls. Calibration error was incorporated to assess how well predicted probabilities aligned with observed clinical outcomes, a critical requirement for risk-stratification systems deployed in healthcare environments. Well-calibrated models ensure that predicted risk scores maintain clinical meaning and support evidence-based decision-making.

To enhance transparency and interpretability, attention-weight analysis and gradient-based attribution methods were used to identify key textual regions influencing the model's predictions. This interpretability assessment helped verify that the model relied on clinically relevant cues rather than spurious artefacts. Finally, k-fold cross-validation was implemented across diverse patient subsets to ensure robustness, reduce overfitting, and demonstrate consistent performance across demographic and clinical variations. This multi-metric evaluation framework ensured that the proposed NLP system not only performed well statistically but also aligned with the stringent expectations of clinical application.

RESULT AND ANALYSIS

4.1 Performance Overview

The proposed NLP framework demonstrated strong predictive performance across all datasets. Compared to traditional machine-learning baselines, the transformer-enhanced architecture achieved higher accuracy in identifying early diabetes risk patterns. The model revealed hidden associations such as recurring respiratory symptoms linked with metabolic stress, subtle mentions of numbness indicating neuropathic development, and familial risk references embedded within consultation narratives.

Table 3: Model Performance Comparison

Model Type	AUC-ROC	F1-Score	Precision	Recall
Logistic Regression	0.78	0.71	0.69	0.73
Random Forest	0.82	0.75	0.76	0.74
ClinicalBERT + BiLSTM	0.91	0.88	0.86	0.90
Proposed Hybrid NLP Model	0.94	0.91	0.90	0.92

The hybrid NLP architecture outperformed baselines by a significant margin, demonstrating the impact of contextual embeddings and temporal text analysis.

4.2 Hidden Pattern Discovery

The system uncovered several risk-enriched linguistic patterns such as:

- Recurrent mentions of fatigue, polyuria, and blurry vision preceding elevated HbA1c readings.
- Medication non-adherence references predicting rapid glycemic deterioration.
- Comorbid mentions such as hypertension and obesity clustering within high-risk groups.
- Implicit lifestyle indicators such as “irregular meals,” “sleep disturbance,” and “reduced activity.”

Table 4: Top Hidden Risk Indicators Extracted from Narrative Text

Indicator Type	Example Extracted Signal	Predictive Significance
Symptom Cluster	"Numbness in feet", "tingling"	Early neuropathy
Behavioural Cue	"Skips medication", "irregular meals"	High progression risk
Emotional Tone	"stress", "low energy"	Metabolic instability
Comorbidity Interaction	"hypertension + obesity"	Multi-risk clustering

4.3 Temporal Progression Analysi

Temporal modelling revealed that symptom clusters evolve in predictable and clinically meaningful sequences, highlighting the value of longitudinal text analysis in diabetes research. The model identified that early references to fatigue, increased thirst, frequent urination, or nonspecific malaise often appeared months before any measurable deviation in HbA1c values, indicating that narrative cues can serve as early indicators of glycemic dysregulation. Patients with repeated mentions of numbness, tingling, or “burning sensations” in lower extremities progressed toward neuropathic or foot-related complications within an average span of two years, supporting the role of temporal NLP in predicting long-term diabetic sequelae. The system further uncovered that subtle behavioural cues such as irregular meal patterns, fluctuating body weight descriptions, reduced physical activity, and episodic sleep disturbances reliably preceded sharp glycaemic fluctuations or medication adjustments.

Temporal attention layers also exposed hidden trajectories in patient behaviour and clinical episodes. For instance, patients who consistently reported emotional stress, anxiety, or work-related fatigue in their follow-up notes exhibited higher likelihood of metabolic instability, demonstrating how psychosocial factors embedded in text contribute to disease progression. Similarly, sequences of missed follow-up visits, inconsistent reporting of medication adherence, and recurring mentions of “feeling better and stopping medicines” were strong predictors of eventual clinical deterioration. Notably, the model captured the interplay between comorbidities such as hypertension and hyperlipidemia, where textual mentions of poorly controlled blood pressure often preceded combined metabolic complications.

CONCLUSION

The study demonstrates that NLP-driven analysis of EHR narratives offers a powerful means of identifying hidden diabetes risk patterns that are not captured by structured clinical data. The proposed hybrid model, integrating transformer embeddings, temporal sequencing, and medical concept extraction, outperformed traditional methods in prediction accuracy, interpretability, and early-risk detection capabilities. By uncovering subtle linguistic clues embedded in physician notes and follow-up narratives, the system enhances the ability to recognise early metabolic deterioration, anticipate complications, and stratify patients more effectively. The findings underscore the transformative potential of NLP in diabetes research, enabling personalised care, proactive interventions, and data-driven clinical support. The research provides a foundational framework for integrating narrative intelligence into chronic disease management, paving the way for more robust, scalable, and clinically meaningful applications of medical NLP.

FUTURE WORK

Future research should leverage larger multi-institutional datasets to improve model generalisability and validate performance across diverse populations. Incorporating speech-to-text medical transcripts, diet logs, and wearable sensor data could deepen multimodal risk identification. Additionally, reinforcement learning architectures may enhance adaptive risk stratification in real-time clinical settings. Further advancement in interpretable NLP methods will be essential for clinical adoption, enabling physicians to understand model-derived insights and integrate them into decision-making workflows. Expanding the system to estimate complication timelines and recommend personalised interventions represents a promising continuation of this work.

REFERENCE LIST

1. Shivade, C. et al., “A review of approaches to identifying patient phenotype cohorts using electronic health records,” *Journal of the American Medical Informatics Association*, 2014.
2. [Luo, Y. et al., “Predicting diabetes risk using electronic health records and machine learning,” *Journal of Biomedical Informatics*, 2016.
3. Esteva, A. et al., “A guide to deep learning in healthcare,” *Nature Medicine*, 2019.
4. Lee, J. et al., “Natural language processing for unstructured EHR data in diabetes research,” *Diabetes Care*, 2020.
5. Devlin, J. et al., “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL-HLT*, 2019.
6. Alsentzer, E. et al., “Publicly available clinical BERT embeddings,” *ACL Clinical NLP Workshop*, 2019.

7. Johnson, A. et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, 2016.
8. Huang, K. et al., "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.
9. Lin, C. et al., "Mining EHR text to identify early diabetic neuropathy," *Journal of Biomedical Informatics*, 2019.
10. Wu, S. et al., "Deep learning for predicting diabetes complications," *IEEE Journal of Biomedical and Health Informatics*, 2020.
11. Lipton, Z. et al., "Learning to diagnose with LSTM recurrent neural networks," *ICLR*, 2016.
12. Rumshisky, A. et al., "Predicting early risk of diabetes using NLP and machine learning," *AMIA Annual Symposium Proceedings*, 2016.
13. Boag, W. et al., "Cliner2: A system for automatically extracting clinical concepts," *AMIA*, 2015.
14. Rajkomar, A. et al., "Scalable deep learning for EHRs," *npj Digital Medicine*, 2018.
15. Jagannatha, A. N., & Yu, H., "Structured prediction models for RNN-based sequence labeling in clinical text," *EMNLP*, 2016.
16. Solares, J. R. A. et al., "Deep learning for electronic health records: A comparative study," *PLOS Digital Health*, 2023.
17. Li, I. et al., "Extraction of diabetes-related risk factors from clinical text: A systematic review," *BMC Medical Informatics and Decision Making*, 2019.
18. Wang, Y. et al., "A clinical NLP framework for phenotyping diabetes using EHR text," *Journal of Biomedical Informatics*, 2020.
19. Yang, X. et al., "Clinical concept embedding for diabetes stratification," *Artificial Intelligence in Medicine*, 2020.
20. Zhang, Y. et al., "BioGPT: Generative pre-trained transformer for biomedical text generation," *Briefings in Bioinformatics*, 2023.
21. Lin, C. et al., "Temporal phenotyping using electronic health records," *Nature Communications*, 2019.
22. Miotto, R. et al., "Deep Patient: Unsupervised representation learning for EHRs," *Scientific Reports*, 2016.
23. Shickel, B. et al., "Deep learning in EHR analysis: A survey," *IEEE Journal of Biomedical and Health Informatics*, 2018.
24. Kadra, A. et al., "Deep learning for detection of undiagnosed diabetes from clinical narratives," *JMIR Medical Informatics*, 2021.
25. Chen, T. et al., "Attention-based neural architectures for clinical text analytics," *ACL BioNLP Workshop*, 2020.