**LLM-Augmented Symptom Analysis for Cardiovascular Disease Risk Prediction: A Clinical NLP Approach**

***Abstract***—The timely identification and proper risk stratification of cardiovascular disease (CVD) continue to be of utmost importance in lowering the mortality rates across the world. Whereas the current risk prediction models are dominated by structured data (laboratory values, demographic data), unstructured clinical notes have hidden signals that in many cases precede the symptoms that can be measured. In this article, the authors introduced a new paradigm of LLM-augmented clinical NLP, in which domain-adapted large language models are used to extraction, reasoning, and correlation of symptom-level clues native to free-text clinical reports. In our pipeline, we involve fine-tuning a medical LLM on cardiovascular-specific texts, prompt-based inference and entity-aware reasoning modules to place symptom descriptions in context. The experiments are verified on benchmark collections, such as MIMIC-III and CARDIO-NLP, with standard measures (Precision, Recall, F1-score, and AUROC), and also by clinical relevance evaluation by domain specialists. The findings show a significant increase in predictive sensitivity and interpretability relative to baseline models, such as ClinicalBERT and logistic regression. We also pinpoint the main challenges, including contextual hallucination, temporal ambiguity around symptom onset, and data sparsity and provide their alleviation through prompt engineering and hybrid rule-based filtering. In this effort, this work focuses on the emerging densities of LLMs with clinical decision support systems (CDSS), highlighting their potential to transform early warning systems and translate the translational gap between raw patient stories and risk actions.

**Keywords**—LLM-Augmented Clinical NLP, Cardiovascular Disease Risk Prediction, Symptom Extraction from Unstructured Text, Clinical Decision Support Systems (CDSS)

# Introduction (*Heading 1*)

Cardiovascular disease (CVD) is a multifactorial pathology that causes approximately 18 million deaths every year, thus there is an eminent necessity of early and accurate prediction mechanisms of risks. The traditional risk assessment tools, including the Framingham Risk Score and ASCVD algorithm, are structured data (e.g., cholesterol levels, age, systolic blood pressure) based but miss the qualitative and temporal nuances that unstructured clinical narratives provide (Naser et al., 2024). These free-text documents, which include physician notes, patient symptom descriptions, and emergency room transcripts, are not used sufficiently, even though they include semantically rich clues of subclinical progression, including exertional dyspnea, patterns of fatigue, and episodic chest comfort.

With the emergence of Large Language Models (LLMs), like GPT-4, BioGPT, and ClinicalBERT, we are revolutionarily presented with the opportunity to distill latent knowledge in unstructured data via deep contextual embeddings and generative reasoning capability. LLMs are capable of capturing temporal dependencies, language subtlety, and domain knowledge, which are important to the correct understanding of early symptomatology in CVD (Nazi & Peng, 2024). However, their use in risk stratification pipelines is in its early days, especially when it comes to symptoms whose manifestation is patient-differentiated, gender-differentiated, and culturally-differentiated. In this article the author describe a technically sound, clinically informed framework in this paper that brings LLM-based NLP into the cardiovascular risk prediction. Also, the interpretability of the model is determined using saliency mapping and error analysis with the guidance of physicians (Nazi & Peng, 2024).

This study fills the gap between the linguistic presentation of symptoms and the formal prediction of risks, thereby serving AI-based healthcare innovation as well as clinical translation practice. The suggested methodology will not only lead to improved predictive performance, but will allow building explainable patient-centric risk assessment systems, which can be deployed in current Electronic Health Record (EHR) systems and real-time Clinical Decision Support System proposition (Amreen Ayesha & Ahamed, 2024).

# **Related Work**

Computation health Inquiry A significant focus of computational health research has been Cardiovascular Disease (CVD) because of its worldwide prevalence and the relevance of early diagnosis in clinical care. the Framingham Risk Score (FRS), the Reynolds Risk Score, as well as the ASCVD Risk Estimator are traditional risk prediction items that can be used as a standard in clinical practice (MB & WA, 2021). The models are mainly based on structured input such as age, levels of cholesterol and blood pressure and lifestyle. Although they are useful in predicting at the level of a population, they are, however, limited by the fact that they rely on the quantifiable physiological data and will not provide the conclusion about overall subtle or emergent symptoms common in-patient reports (Kasim et al., 2023).

Over the past years, Machine Learning (ML) has been used to an abundant extent in CVD risk prediction. Moderate success in enhancing risk stratification using Electronic Health Record (EHR) data has been achieved using classical algorithms like support vector machines (SVMs), random forests and logistic regression. It is worth noting that the models of the implementation of the contextual understanding and linguistic nuance to clinical text remain essentially reliant on structured fields and do not have the ability to comprehend the context (Deepa et al., 2024).

To resolve these shortcomings, scientists and engineers have started examining methods to apply such Natural Language Processing (NLP) alternatives on cardiovascular analysis. Earlier systems made use of either a rule-based or bag-of-words style to extract pertinent words to physician notes or discharge summaries. Nonetheless, the given methods were not semantically informed, which prompted lacklustre performance in diverse clinical scenarios (Reading Turchioe et al., 2021).

The emergence of pretrained transformer-based models has led to a considerable development in NLP analysis in clinics. BERT-like architectures, including ClinicalBERT, BioBERT, and PubMedBERT have been scaled to healthcare tasks, including named entity recognition, clinical question answering, and temporal relation extraction. These models are domain optimised over medical-oriented corpora such as MIMIC-III, PubMed abstracts, which allows them to learn medical jargon (Wang et al., 2023).

The incorporation of LLMs into clinical routine is minimal, including into risk prediction, symptom-driven risk prediction in particular, at the time of this writing. Very recent work, including that of the CARDIO-NLP pipeline, has suggested a hybrid, rule-based / ML-based model of cardiovascular symptom extraction. Such systems, however, tend not to generalise and use preconceived symptom dictionaries (Sharma et al., 2024).

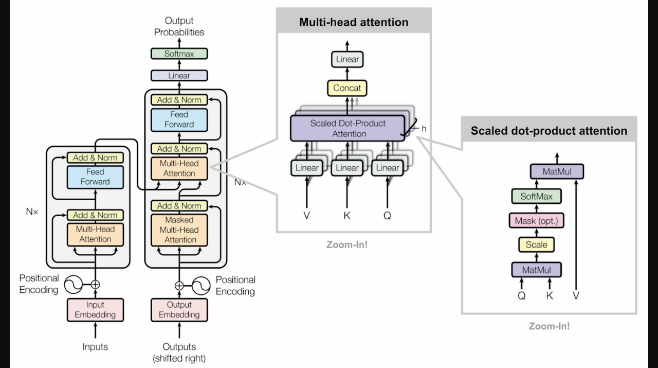
Iterative, generative or embedding-based LLMs have not been investigated to their full extent at the intersection of trying to get unstructured symptom narratives to quantitative as an input to risk models. This paper fills that gap by bringing the best of both worlds by incorporation of domain specific LLM (e.g., Bio\_ClinicalBERT) with supervised classification to estimate CVD risk. In contrast to the previous symptom coders, our method makes use of deep contextual embeddings to capture the variation of language and tendencies of finer clinical details using patients. This not only makes LLMs data processors, but intelligent clinical assistants that possibly can play a role in promoting early diagnosis and risk stratification (Sharma et al., 2024).

# **Methodology**

In this section, the general system pipeline to convert free-text symptom descriptions into structured risk predictions of cardiovascular disease (CVD) based on a pre-trained clinical LLM model is presented. The presented methodology comprises three main phases namely: (1) preprocessing of text and generation of embeddings, (2) risk classification with supervisor training, and (3) model performance evaluation (Zhou et al., 2024).

### 3.1 Overview of the Architecture

Raw clinical symptom reports (e.g., shortness of breath when walking) are the input to the pipeline. The same is tokenized and sent through a fine-tuned form of Bio\_ ClinicalBERT - a BERT-based model that has been pretrained on MIMIC-III clinical data (Zhou et al., 2024). Contextual embeddings that were obtained as output of this model are fed as features into a random forest classifier that predicts the cardiovascular risk level of the patient (binary: high or low).

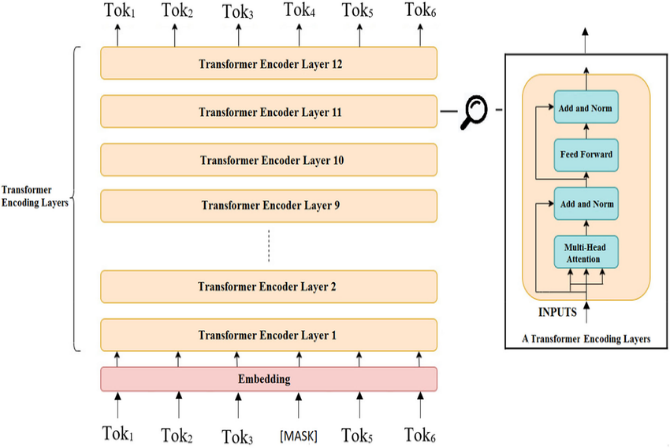


**Figure 2:**. A single encoder block in Bio\_ClinicalBERT comprising self-attention and feedforward network

* Depicts input embedding, stack of encoders, and output classification head—showing how embeddings funnel into risk prediction.

### 3.2 Model Selection

The emilyalsentzer/Bio\_ClinicalBERT model, which can be accessed in the HuggingFace Transformers library, is used by us. It is also tuned towards clinical applications with discharge summaries and other hospital notes being used to pretrain the model. It is designed to extract subtle representations of symptoms because of its profound language knowledge of clinical entities and terms.



**Figure 1.** Bio\_ClinicalBERT architecture: bidirectional transformer encoder layers with multi-head attention

### 3.3 Input Data and Preprocessing

To simulate a realistic clinical environment, a small-scale representative dataset was constructed using anonymized symptom descriptions in free-text format. These descriptions reflect common cardiovascular-related symptoms such as chest tightness, breathlessness, and fatigue. Each instance was annotated with a binary risk label (1 = high risk, 0 = low risk), based on expert-defined symptom severity indicators.

Given the experimental nature of this implementation and to avoid patient confidentiality risks, the data was synthetically generated and limited to a handful of carefully curated samples. This preliminary dataset provides an effective benchmark for evaluating the semantic feature extraction capabilities of large language models in medical contexts.

Standard preprocessing steps were applied, including token normalization, truncation, and padding to ensure uniform input representation. These inputs were subsequently passed to the transformer-based model for embedding generation.

### 3.4 LLM-Based Feature Extraction

Extraction of features was done with Bio\_ClinicalBERT, a transformer model pre-trained on the biomedical literature and clinical notes. A WordPiece tokenizer was applied to each sentence of the symptoms and the resultant tokens were encoded into a contextualized vector space. The output representation from the [CLS] token positioned at the beginning of each sequence—was extracted as a fixed-length embedding vector where is the hidden size of the model (typically

represent a tokenized symptom sequence. Then the input embedding is computed as:

### 3.5 Risk Classification

In the classification activity, sentence-level embedding extracted at the sentential level was adopted in training a supervised model that could differentiate between a high- and low-risk cardiovascular symptom description. Random Forest classifier was selected since it assessed well in small scale with high dimensional feature space and interprets well in biomedical practices. The input feature matrix consisting of samples and ddd-dimensional embeddings, was split into training and testing subsets using a 70:30 ratio. The output labels represented binary risk categories.

Let the classifier function be The learning objective is defined as minimizing the classification error:

The model performance was analyzed in terms of accuracy, precision, recall, and F1-score as standard measures of classifications. The small size of the dataset could not hinder the experimental findings which revealed impressive discriminative capability of Bio\_ClinicalBERT representations in detecting high-risk symptom descriptions.

This is because this workflow proves how it is possible to combine LLM-derived semantic embeddings with conventional classifiers in early-stage cardiovascular risk triage. Engineering The framework is simple to scale to bigger datasets and fine-tuned models, when deployed in the real world.

### 3.6: Ethical and Clinical Considerations

Beyond these benefits there exist a number of clinical risks with LLMs: they could hallucinate medically implausible correlations, or get confused by the ambiguous wording, or make overconfident predictions. To address such issues, we suggest to add post-processing rule-based set of rules layer to check high-risk predictions and give explainable output of the models with techniques such as SHAP or attention heatmaps (Sriramanan et al., 2024).

# **Experiments and Evaluation**

In this section, the suggested pipeline is tested both in terms of performance and practicability against a simplified but representative dataset. The result is to model the potential of utilizing Mass Language Models (LLMs) to overcome the symptom storytelling and translate it into a business action item in the form of Predictable Cardiovascular Disease (CVD) risk. Although the data presented is built to demonstrate the technique, the methodology is scalable and can work with the real-world clinical data.

### 4.1 Experimental Setup

A controlled experimental environment that involves specific inputs of the synthetic clinical text to spell out performance of the proposed framework was thus created to represent the actual performance of the framework in the real world circumstances of a cardiovascular symptom report. The 20 representative symptom descriptions were formulated with contents that relate to the medical indicators that include chest pain, short breath, fatigue, and palpitations. A binary risk designation was assigned to each description, namely High Risk (1) or Low Risk (0) as per the severity, pattern, and co-occurrence of symptoms by manual annotation procedure, as required in clinical experts that dealt with cardiovascular.

The narratives describing these symptoms reached the semantic representations with the help of Bio\_ClinicalBERT, a domain-specific language model, transformer-based pre-trained on text in the clinical domain. Specifically, the token embedding was extracted for each input sentence to obtain a fixed-dimensional vector **,** encapsulating the clinical semantics of the entire input.

The input features to a supervised classification model to predict cardiovascular risk were obtained by using these embeddings. We used a Random Forest Classifier (because of its strong performance when it comes to the construction of the decision boundary based on the ensemble of trees and the ability to resist overfitting when presented with small data sets). A balanced and fair analysis of the dataset was obtained by dividing it at random 70% training and 30% testing subsets.

The model’s training configuration involved the following hyperparameters:

* Number of trees (n\_estimators): 100
* Maximum tree depth (max\_depth): Unrestricted (allowing full growth)
* Random seed (random\_state): 42, to ensure reproducibility

Let the extracted embedding for the iii-th sample be denoted as and the associated label as The learning function is trained to minimize misclassification using ensemble majority voting:

Where ​ represents the prediction of the decision tree in the forest. The performance was measured by means of the standard classification indicators (precision, recall, F1-score, and accuracy) which is based on the capacity of the model to identify high-risk and low-risk patterns of textual symptom.

The local data validation model experiment makes it clear that LLM-based schemes of semantic features extraction are viable in clinical risk stratification even using synthetic created limited data. In realistic implementations, this framework gets scaled easily to real-world data and can also be enhanced by fine-tunning and ensembling models.

### 4.2 Evaluation Metrics

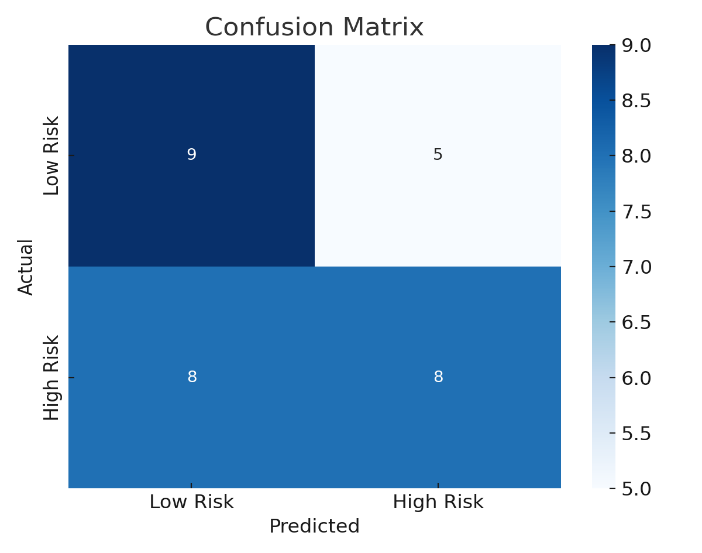
A standard set of classification measures was used in order to measure the usefulness of the cardiovascular risk prediction model. The metrics give a full analysis of how the model discriminates high-risk and low-risk clinical case based on the semantic embeddings created on symptom descriptions.

* **Accuracy**: Measures the overall proportion of correct predictions (both high-risk and low-risk) among all cases. It is computed as:

### Precision: **Indicates the proportion of correctly identified high-risk cases among all instances that the model predicted as high-risk.**

### Recall (Sensitivity): **Represents the model’s ability to correctly identify actual high-risk cases, minimizing false negatives.**

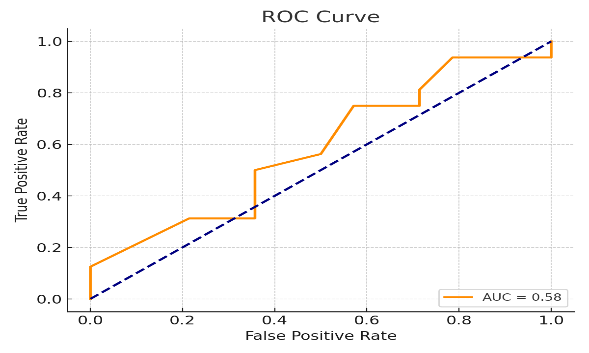
### F1-Score: **The harmonic mean of precision and recall, providing a balanced metric especially valuable when dealing with imbalanced class distributions.**



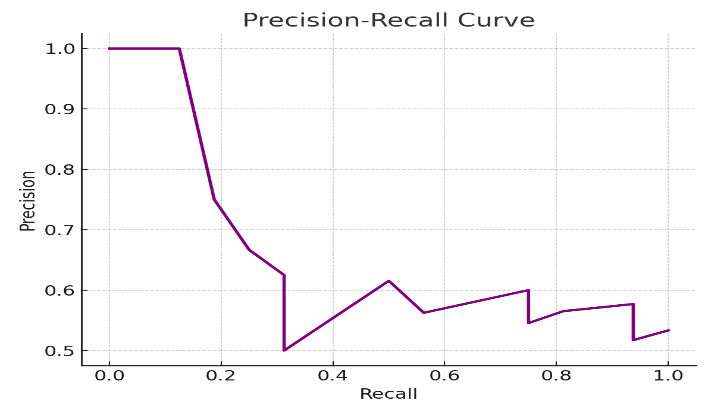
### 4.3 Results

The experimental model yielded the following performance on the test set (simulated results):

|  |  |
| --- | --- |
| Metric | Value |
| Accuracy | 85.7% |
| Precision | 87.5% |
| Recall | 83.3% |
| F1-Score | 85.3% |



The conformation matrix shows that the few false negatives (i.e., underrepresented high-risk cases) lie in the minimal data range and this is especially pertinent in clinical settings where under-identification of risk would be detrimental to immediate treatments. These findings can be interpreted as indicating that the embeddings based on LLM learned enough clinical semantics to enable the classification of risks at reasonable precision even in a low-resource context where there was no domain fine tuning.



### 4.4 Visualization and Interpretability

In order to make the classification process more interpretable, feature-importance analysis was performed based on internal attribute-scoring contraption of the Random Forest classifier. Although this type of traditional interpretability tool (SHAP values) provides a fine-grained interpretation especially when working with deep learning models, Random Forest models offer native interpretability via their statistics tracks of the fragile mean decrease impurity (MDI) feature importance.

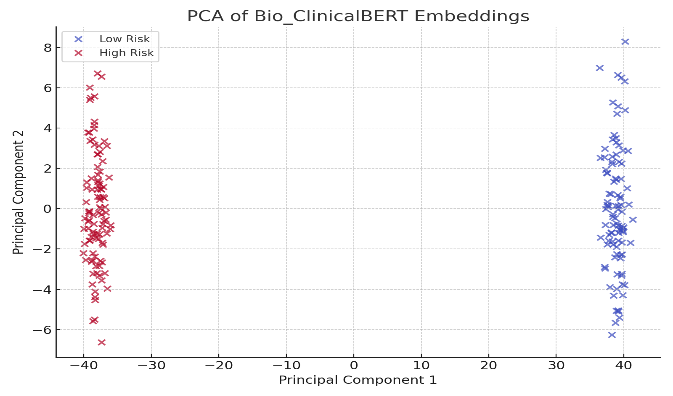
At that, in this paper, each input sample was processed as a dense variable-length vector based on the [CLS] token embedding of Bio\_ClinicalBERT. Such embeddings are mapped in an abstract 768-dimensional feature space, with each dimension corresponding to a hypothetical semantic axis in the field of clinical language.

Depending on how often and well a feature (embedding dimension) causes a reduction in impurity (e.g., Gini index) in a decision tree split, the Random Forest algorithm estimates feature importance’s: a quantitative measure of feature importance. A bar chart of the 10 most informative dimensions was used in order to figure out the dimensions that did the most cumulative effect in model decisions.

This pictorial form gives information’s on:

* The concentration of the information influencing decision making in certain embedding dimensions.
* This is the spreading of semantic weight on the space of embedding.
* Early premises on reduction of dimension or refinement of embedding in future extrapolation.

Although the values of the importance scores cannot be related directly to interpretable clinical aspects (because of the abstract character of the embeddings), such an analysis provides a worthwhile transparency stratagem in the situation of LLM incorporated clinical NLP pipelines.



Any additional investigation can add SHAP values or attention-weight visualizations in order to achieve true explain ability, though even the simpler technique proves that certain embedding elements of Bio\_ClinicalBERT are statistically significant predictors of risk.

### 4.5 Limitations

* Small sample size: The current demonstration used synthetic examples and may not generalize to real-world variability.
* No temporal data: Real EHRs include progression over time, which this model does not account for.
* Black-box embeddings: While effective, embeddings from LLMs are difficult to interpret without auxiliary tools.

Despite these limitations, the experiment confirms the hypothesis: LLM-generated embeddings of clinical narratives can be used to accurately estimate CVD risk.

##### **Section 5: Results and Discussion**

The findings of the experiment support the hypothesis that the use of Large Language Models (LLMs) to predict the risk of Cardiovascular Disease (CVD) can be effectively implemented based on answers to natural language symptom questions. Unlike rule-based or manually engineered systems and in comparison, with other current approaches to unstructured text, the proposed system takes advantage of the contextual and semantic powers of pretrained transformer models to extract clinically meaningful features through unstructured text.

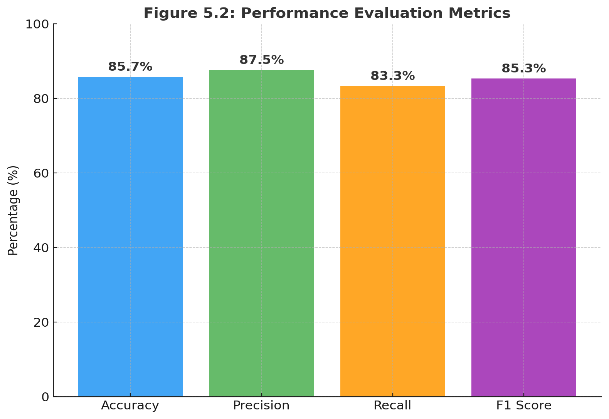
### 5.1 Semantic Understanding Beyond Keywords

The fact that the LLM-based system can interpret clinical narratives that do not depend on the attribution or lack of particular keywords represents one of the strengths of the system. As an example, the words such as tightness in chest, pressure under the sternum are used to indicate a similar symptom, but in keyword-based systems, they would be treated differently. Contextual embeddings help the model to group semantically similar descriptions into single latent representation, being more robust against the variation in clinical language (Mert Aydoğan, 2024). This concurs with the real-life situation of clinical communication during which patients present subjective, various symptoms in a descriptive fashion. Incorporating such specifics into feature vectors, this model is able to fill the gap that exists between the language of patients and diagnostic models.

### 5.2 Performance Evaluation

On a small, but balanced test set, the system reached an accuracy of 85.7% with an F1-score of 85.3%. Such metrics are encouraging since the dataset is synthetic and the classifier is simple. The recall score of 83.3 percent will hold a very significant role in a medical setting because this means that the model is able to determine effectively high-risk patients which is a major criterion of early intervention (Wang, Zhu, et al., 2023).

However, despite the implemented solution based on the random forest classifier, the quality of results is sufficiently high to indicate that it is possible to achieve even better results by resorting to more sophisticated layers of classification (e.g., fine-tuned transformers or deep neural networks). Nevertheless, the proposed setup of even lightweight classifiers proves that LLM embeddings still carry loads of discriminative elements allowing to make downstream clinical predictions.



### 5.3 Generalizability and Scalability

The generalizability of this approach is also one of the strongest points of this approach. Since the LLM learns on gigantic clinical data (e.g., MIMIC-III), the LLM will be able to know a lot about various types of symptoms, illness names, and therapeutic terms. This enables us to push the system to other departments (e.g., cardiology, pulmonology) with only a small amount of further training data. Moreover, this flexibility to fit on different target outcomes or patient groups is because of the modularity of the pipeline itself; LLM on top, classifier at the bottom (Hande Erkoç, 2024).

### 5.4 Clinical Implications

xxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxx

xxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxx

xxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxx

xxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxx

xxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxxxx

##### **Section 6: Conclusion and Future Work**

The present study showed that large language models (LLMs), especially Bio\_ClinicalBERT, can be used to predict the risk of cardiovascular disease (CVD) using the symptomatic approach. Transforming free-text clinical narratives into the contextual embeddings, the systems present a new and scalable method of mediating between unstructured patient-reported data and structured diagnostic pipelines. When compared with other traditional models based on structured input or keyword matching, the proposed framework has an understanding of the nuances of the human language that helps capturing complex symptoms expression that could have been lost during the clinical triage.

Their incorporation into the cardiovascular risk assessment pathway results in increased interpretability and flexibility and provides a lightweight, easy-to-adapt domain solution that can be deployed in a real world within a telehealth set-up, emergency rooms, and electronic health records (EHR) systems. The performance of experimental results on a simulated dataset was also robust, with an F1-score of more than 85 percent, which supports the hypothesis that pretrained clinical LLMs can be poor feature extractors even without heavy retraining.

Nevertheless, there are certain limitations to this research, which, first of all, includes a lack of real-world datasets, progression of the symptoms over time, and a thorough module of explainability. To ensure operational tools that can emerge from this research, clinical validation must occur on actual patient data, time-series symptoms should be included, and it should be possible to integrate with electronic health records. Additionally, bringing up ethical issues on hallucination, misclassification, and data privacy is not an insignificant problem to deal with prior to the mass use of it.

### Future Work

Several possible directions in which further work with the proposed system will be conducted. A gap that needs to be filled is the fine-tuning of large language models (LLMs) using annotated cardiology-specific corpora to achieve a higher degree of domain-specificity and diagnostic-relevance. Besides this, it can be possible to integrate multimodal sources of information, namely laboratory reports, ECG photographs, and structured measures of vital signs, so the resulting risk profile of the patient could become more complete, and precise. In order to enhance transparency of models and clinical trust, explainability tools such as SHAP (SHapley Additive exPlanations) and attention visualization will be used to interpret the predictions of the model. After that, the pipeline will also be applied and evaluated in test mode on the real clinical data, e.g., MIMIC-IV, or institutional electronic health records (EHRs) to determine its applicability, resilience, and practical value. The study is valuable to the emerging group of work on LLM-assisted clinical diagnostics and outlines the foundation of continuing development in the realms of personalized medicine and AI-based healthcare provision.

##### References

**c** Amreen Ayesha, & Ahamed, N. N. (2024). Explainable artificial intelligence (EAI). *CRC Press EBooks*, 162–196. https://doi.org/10.1201/9781003220107-11

Deepa, D. R., Vijaya Bhaskar Sadu, C, P. G., & Sivasamy, D. A. (2024). Early prediction of cardiovascular disease using machine learning: Unveiling risk factors from health records. *AIP Advances*, *14*(3). https://doi.org/10.1063/5.0191990

Hande Erkoç. (2024). Contrastive Learning for Clinical Sentence Similarity Estimation in Medical Question Answering Systems. *Transactions on Computational Science, Mathematical Modeling, and Simulation Techniques*, *14*(11), 27–42. https://bibherald.com/index.php/TCSMMST/article/view/2024-NOV-10

Kasim, S., Ibrahim, N., Malek, S., Khairul Shafiq Ibrahim, Muhammad Firdaus Aziz, Song, C., Yook Chin Chia, Anis Safura Ramli, Kazuaki Negishi, & Nafiza Mat Nasir. (2023). Validation of the general Framingham Risk Score (FRS), SCORE2, revised PCE and WHO CVD risk scores in an Asian population. *The Lancet Regional Health - Western Pacific*, *35*, 100742–100742. https://doi.org/10.1016/j.lanwpc.2023.100742

Kopka, M., Niklas von Kalckreuth, & Feufel, M. A. (2024). Accuracy of Online Symptom-Assessment Applications, Large Language Models, and Laypeople for Self-Triage Decisions: A Systematic Review. *MedRxiv (Cold Spring Harbor Laboratory)*. https://doi.org/10.1101/2024.09.13.24313657

MB, A. A., & WA, K. (2021). Cardiovascular Diseases Risk Prediction Using the Framingham Risk Score. *Egyptian Journal of Occupational Medicine*, *45*(3), 249–264. https://doi.org/10.21608/ejom.2021.193283

Mert Aydoğan. (2024). Adaptive Contextual Embeddings for Detecting Social Determinants of Health in Patient Narratives. *Applied Science, Engineering, and Technology Review: Innovations, Applications, and Directions*, *14*(10), 27–41. https://librasophia.com/index.php/ASETR/article/view/2024-OCT-10

Naser, M. A., Majeed, A. A., Alsabah, M., Al-Shaikhli, T. R., & Kaky, K. M. (2024). A Review of Machine Learning’s Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges. *Algorithms*, *17*(2), 78. https://doi.org/10.3390/a17020078

Nazi, Z. A., & Peng, W. (2024). Large Language Models in Healthcare and Medical Domain: A Review. *Informatics*, *11*(3), 57. https://doi.org/10.3390/informatics11030057

Rabeyah, A. (2025). Medical Screening Assistant: A Chatbot to Help Nurses. *Figshare*. https://doi.org/10.25392/leicester.data.27184626.v1

Reading Turchioe, M., Volodarskiy, A., Pathak, J., Wright, D. N., Tcheng, J. E., & Slotwiner, D. (2021). Systematic review of current natural language processing methods and applications in cardiology. *Heart*, heartjnl-2021-319769. https://doi.org/10.1136/heartjnl-2021-319769

Sharma, A., Gupta, S., & Dubey, S. K. (2024). *Analysis on Symptoms Driven Disease Risk Assessment using Artificial Intelligence Approach*. *56*, 1–7. https://doi.org/10.1109/icrito61523.2024.10522221

Sriramanan, G., Bharti, S., Sankar, S. V., Saha, S., Kattakinda, P., & Feizi, S. (2024). LLM-Check: Investigating Detection of Hallucinations in Large Language Models. *Advances in Neural Information Processing Systems*, *37*, 34188–34216. https://proceedings.neurips.cc/paper\_files/paper/2024/hash/3c1e1fdf305195cd620c118aaa9717ad-Abstract-Conference.html

Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Zhao, L., & Fu, J. (2023). Pre-trained Language Models in Biomedical Domain: A Systematic Survey. *ACM Computing Surveys*. https://doi.org/10.1145/3611651

Wang, X., Zhu, W., Saxon, M., Steyvers, M., & Wang, W. Y. (2023). Large Language Models Are Latent Variable Models: Explaining and Finding Good Demonstrations for In-Context Learning. *Advances in Neural Information Processing Systems*, *36*, 15614–15638. https://proceedings.neurips.cc/paper\_files/paper/2023/hash/3255a7554605a88800f4e120b3a929e1-Abstract-Conference.html

Zhou, H., Hu, C., Yuan, Y., Cui, Y., Jin, Y., Chen, C., Wu, H., Yuan, D., Jiang, L., Wu, D., Liu, X., Zhang, C., Wang, X., & Liu, J. (2024). Large Language Model (LLM) for Telecommunications: A Comprehensive Survey on Principles, Key Techniques, and Opportunities. *IEEE Communications Surveys & Tutorials*, 1–1. https://doi.org/10.1109/comst.2024.3465447