

NLP-XAI Based Model in Early Diagnosis of Cardiovascular Diseases: A Comprehensive Review

Yatindra Pathak

Research Scholar

FET, Rama University Kanpur, UP
yp02.pathak@gmail.com

Dr. Somendra Tripathi

Associate Professor

FET, Rama University Kanpur, UP
somendra.fet@ramauniversity.ac.in

Abstract— cardiovascular diseases consists a major factor in global healthcare nowadays and global loss of life and the cost burden on global healthcare can be reduced by well-timed recognition of cardiovascular diseases. In today's world a massive use of electronic health records also produces a large amount of unstructured medical data. This data can lead us to precedent symptoms factor and course of disease. In this context natural language processing becomes an enabling technology for predicting future insights in this unstructured data. Cardiovascular diseases produces a major challenge for healthcare system today and early detection of these diseases becomes a crucial need for refining patient outcomes. NLP based systems can easily find the hidden symptoms from this kind of electronic health records. Machine learning and deep learning is widely used for developing clinical decision making systems. Such a system helps medical professionals to identify cardiovascular disease more precisely and accurately. In this review we discussed in detail about early detection of cardiovascular disease using NLP-based model, datasets, evaluation, transform based architecture for deep learning model, standard machine learning techniques and comparative performance. Integration with explainable AI, practical clinical implementation issues, constraints and future scoops are also examined.

Keywords— *Cardiovascular Diseases, Clinical Decision Support, Deep Learning, Early Diagnosis, Electronic Health Records, Machine Learning, Natural Language Processing.*

INTRODUCTION

As per the records provided by the World health organization (WHO), each year more than 17% of worldwide deaths are due to cardiovascular diseases. It is likely 9 million death worldwide are associated with them. That's why it is one of the most considerable and persistent global health burden [1]. However we already moved forward and get advanced in precautionary cardiology, medicine therapies and diagnostic imaging technology. But prior and accurate identification of cardiovascular risk factor is still a significant problem. Versatile nature of patient data and dissimilarity of physician documentation style makes a major impact on clinical decision making progress. Recent healthcare system consist large amount of clinical data in unstructured format i.e. hospital discharge records, electrocardiogram ECG reports, Eco cardiography reports, angiography result and doctor's notes. This unstructured data hides contextual facts such as symptoms improvements, medication requirements and lifetime indicators, which can be identified and used for comprehensive cardiovascular risk detection [2]. it is very difficult to explored this textual data in Conventional computational methodologies and rule based classifications,

due to the limitation of language interpretation and contextual reasoning. Natural language processing or NLP has the capability to make it structured and find out the machine readable insights in it. Due to which it becomes an important tool in biomedical informatics [3]. To extract the clinical insights interrelated to cardiovascular health and their semantic associations, topical transformer based natural language processing architectures like BERT, BioBERT and clinical BERT can be used. However, these methods are very rich in performance to get insights from unstructured data, dekh typically operates as black boxes. And user or doctors cannot identify the reason behind the conclusions which are made by these models. This deficiency of interpretability leads to a serious question about trustworthiness, responsibility and ethical application.

High anticipated accuracy and transparent comprehensible reasoning both have to be achieved by artificial intelligence model in the healthcare system, because patient survival can be directly affected by conclusions taken by the system. Existing deep learning methods hardly provide the reason behind specific risk estimation for cardiovascular studies i.e. which symptoms pointed to the current result, bio markers and textual appearances [4]. This lack of transparency in interpretability works as a stumbling block to real world recognition and physician acceptance of the method. It is necessary to develop and explainable artificial intelligence integrated natural language processing framework that can perform domain appear cardiovascular disease diagnosis and risk estimation as well as provide the reason behind this estimation and diagnosis. This framework is capable of reducing the gap among model's intelligence and clinical reasoning by providing human understandable explanations and leads to expanding the safety, trust and clinical usability of AI driven cardiovascular analytics.

LITERATURE REVIEW

Automated The black box nature of the deep learning models is becoming a questionable mark on using these models in clinical informatics, and explainable artificial intelligent (XAI) can become the solution of this question [5]. Decisions suggested by a model can directly influence patient outcomes in high-risk fields such as cardiology and that's why explain ability is not only a technical but also a behavioral and legal requirement. Early applications of AI in cardiovascular research using convolutional neural networks and recurrent neural networks for arrhythmia detection, myocardial infarction prediction, and heart failure diagnosis,

primarily concentrated on signal-based data, such as electrocardiograms and echocardiography[6][7]. Although these models outperformed traditional statistical methods in terms of accuracy, they also offered very little insight into the decision-making process, which reduced clinician confidence and slowed clinical adoption.

Detailed information on symptoms, diagnosis, prescriptions, and doctor evaluations contain unstructured text data in Electronic Health Records (EHRs) and attracted attention due to the increasing amount of healthcare data. This increases the interest in application of Natural Language Processing (NLP) to identify patterns that are clinically meaningful for future prediction. Recent developments in deep language models such as BERT, BioBERT, and ClinicalBERT have been found very accurate for tasks involving disease categorization and biomedical text mining [8][9]. These models use the transformer architecture that Vaswani et al. introduced. Reference [5] and developed a method that improves feature representation and supports bidirectional contextual comprehension. Even with their accuracy, these models remain in black box category (i.e. offering limited understanding of how clinical terminology and language signals impact final predictions).

LIME [10], SHAP [11], and Grad-CAM [12] like Post-hoc analyzability methods have been also checked for improving transparency in healthcare AI. Facts or characteristics that influence model results are focused by these techniques, but they sometimes fail to concentrate on the Gordian mental processes that clinicians use when making decisions, and also they are Backend-agnostic. For case in point, SHAP offers comprehensive feature importance but is incapable to put on view contextual connections in chronological text, whereas LIME generates explanations for each sample devoid of adopting domain-specific doctrine information. There may be a disagreement between algorithmic and medical accountability if these explanations do not track medically acknowledged lines of facts.

Topical study initiatives have paid attention to developing domain-adapted XAI architectures that go further than post-hoc fancying. Transform based architectures take account of techniques such as gradient attribution and attention-weight visualization to represent which textual elements influence model predictions. By producing token-level explanations and concerning them to clinical terminology, SIDU-TXT [13] eclipses as an vital development that overcomes the gap flanked by AI-generated reasoning and human interpretability. Researchers like Nguyen et al.[14] and Kim et al. [15] address the magnitude of hybrid explainability approaches that take in both quantitative consideration and qualitative participation from medical professionals. By using physician understanding to authorize AI interpretations, such human-centered approaches undertake salutary significance and security.

Explicable healthcare AI is becoming more widespread, but there is still modest study that explicitly uses XAI for cardiovascular natural language processing and the preponderance of cardiovascular AI follow a line of investigation, which focuses on controlled data like ECGs, vital signs or lab examination but still places a privileged

priority on predicting accurateness than interpretability[16][17]. There have been hardly any attempts to apply interpretable NLP pipelines to medical text for identifying risk factors, analyzing symptoms, or prognosis treatment outcome. Key interest for innovation in this less studied field is to Designing frameworks that pull off state-of-the-art performance while providing apparent and clinically reasoned explanations. By improving model liability, refining clinician confidence and allowing for moral AI integration into genuine healthcare systems, the development of a XAI-NLP model intended to identify cardiovascular disease could tackle a gap in the literature.

Table 1 An overview on major research in XAI and NLP integration in Health Care

Author (s) & Ref. No.	Year	Methodology / Model Used	Key Findings	Limitations
Doshi-Velez & Kim [3]	2017	Conceptual framework for interpretable AI	Established theoretical foundation for XAI	No empirical cardiovascular application
Madsen et al. [4]	2022	Evaluation of LIME/SHAP in Healthcare	Highlighted failures of post-hoc tools to ensure trust	Lack of domain-specific Adaptation
Vaswani et al. [6]	2017	Transformer/ Attention Mechanism	Foundation for BERT, BioBERT, ClinicalBERT	No direct clinical focus
Lee et al. [8]	2020	BioBERT	Improved biomedical entity extraction	Limited interpretability
Alsentzer et al. [9]	2019	ClinicalBERT	Enhanced contextual understanding in EHRs	Operates as a black box
Ribeiro et al. [10]	2016	LIME	Introduced local explanation method	Unstable on complex models
Lundberg &	2017	SHAP	Quantified global	

10.48047/jocaaa.2024.33.08.475

Lee [11]			feature importance	Context-insensitive for text
Selvaraju et al. [12]	2017	Grad-CAM	Useful for image interpretation	Limited text applicability
Jahromi et al. [13]	2024	SIDU-TXT	Generated clinically aligned explanations	Early-stage; limited validation
Nguyen et al. [14]	2023	Hybrid CNN-XAI	Improved interpretability in heart disease detection	Restricted to structured signal Data
Kim et al. [15]	2024	Systematic Review	Emphasized need for clinician validation	Did not address cardiovascular text
Johnson et al. [16]	2016	MIMIC-III	Public dataset for cardiovascular AI	Requires preprocessing for NLP
Wagner et al. [17]	2020	PTB-XL	Supports deep learning benchmarking	No text modality
Dey et al. [18]	2023	Deep learning on HER	Demonstrated AI potential for risk prediction	Lacked transparency and Explainability

The rise in research publications on XAI or NLP in medical diagnostics is another important factor to consider. The following graph shows projected publishing statistics based from 2015 to early 2025 based on research trends in Artificial Intelligence, NLP (Natural Language Processing) and XAI (Explainable AI) in medical diagnosis.

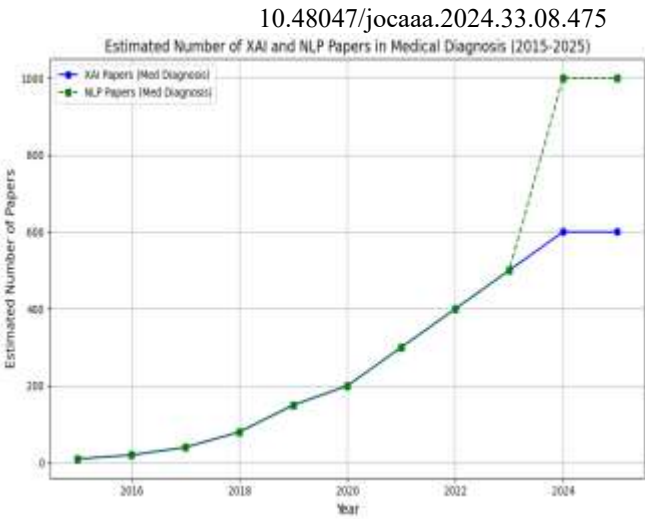


Figure 1 XAI and NLP research trends in medical diagnostics

METHODOLOGY

The framework demonstrates how medical textual data are processed through an NLP model to generate diagnostic or risk predictions. Alongside prediction, an XAI module explains how the result is derived by identifying key words and medical factors influencing the decision.

By combining prediction with explanation, the system provides understandable and trustworthy results, improving transparency and supporting clinician-centered decision-making.

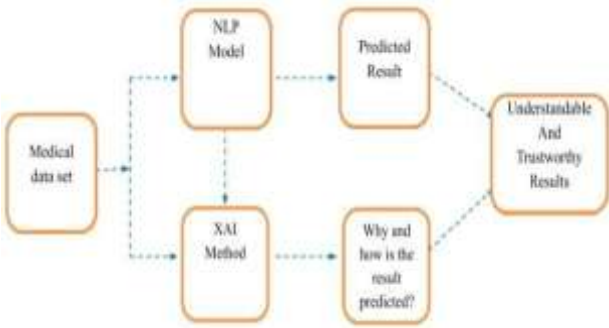


Figure 2 Overview of Framework XAI-NLP Integration over Medical Dataset

Architecture Framework and Design: Recent research recommends slot in Explainable Artificial Intelligence (XAI) into transformer-based medical NLP frameworks to report the explainability boundaries of NLP models in cardiovascular risk-based clinical decision-making. While miniature versions help putting into practice on mobile devices [21] and in customary hospital arrangements, these XAI-integrated models improve transparency, unimpeachable, and constancy. Below is a instant of the generalized framework.

		True value	
		T (Positive)	N(Negative)
Estimated Values	T (Positive)	TP(True Positive)	FP(False Positive)
	N(Negative)	FN (False negative)	TN(True Negative)

F1-score, accuracy, recall, precision, and Matthews Correlation Coefficient (MCC) [16][20].

Table 2 Confusion matrix

The Performance of classification problem can be stated by a matrix of confusion, similar to that shown in Table 2.

A negative labeled (i.e. $y_i=0$) designates the majority class, while a positive labeled (i.e. $y_i=1$) designates minority class.

Precision (P), Recall (R), False Positive (FP), False Negative (FN), and F1 Score served as the foundational measures for assessment.

Several crucial metrics are derived by using the confusion matrix.

TP(True Positive): Actual positive class is anticipated precisely as the positive class.

FP(False Positive): Actual negative class is anticipated imprecisely as positive class.

TN(True Negative): Actual negative class is anticipated precisely as negative class.

FN (False negative): Actual positive class is anticipated imprecisely as negative class.

$$\text{Precision} = \frac{\text{True Positive}}{\text{Total Positives}} = \frac{TP}{(TP + FP)} \quad \dots\dots (1)$$

$$\begin{aligned} \text{Recall/Sensitivity/ Hit - Rate} &= \frac{\text{True Positive}}{(\text{True Positive} + \text{False Negative})} \\ &= \frac{TP}{(TP + FN)} \quad \dots\dots\dots (2) \end{aligned}$$

$$\text{F1 - score} = \frac{2\text{PrecisionRecall}}{(\text{Precision} + \text{Recall})} \quad \dots\dots\dots (3)$$

$$\begin{aligned} \text{Accuracy} &= \frac{(\text{True Positive} + \text{True Negative})}{\text{Total Values}} \\ &= \frac{TP + TN}{(TP + FP + TN + FN)} \quad \dots\dots\dots (4) \end{aligned}$$



Figure 3 Data analysis Architecture for Cardiovascular Text Analysis

Phase 1 Data Acquisition consists PhysioNet [18], PTB-XL[17], and MIMIC-III [16] or likely publicly available sources are used to gather cardiovascular clinical text data. Phase 2 Data Preprocessing is required To safeguard semantic consistency, preprocessing consists of tokenization, stop-word deduction, lemmatization, normalization, and medical term standardization by means of UMLS and SNOMED CT. Phase 3 Model Design introduced that BioBERT, ClinicalBERT, and MedRoBERTa are transformer-based models practiced for risk prediction and cardiovascular entity mining. Lightweight versions permit for resource-efficient deployment and field adaptation improves accepting of vocabulary exclusive to cardiology. Phase 4 deals with Explainability Integration Textual heatmaps and word- significance scores are shaped using XAI tactics such as LIME, SHAP, and SIDU-TXT. To ensure therapeutic relevance, explanations are corroborated against professional annotations. Phase 5 System Deployment & Evaluation includes Measuring models by the help of Accuracy, F1-score, ROC-AUC, MCC etc. Interpretability is inspected through qualitative grading methods and clinician feedback form [19]. After that we need System Deployment that includes a dashboard visible to clinicians that provides diagnostic predictions and detailed explanations. This setup encourages open communication between humans and AI during the assessment of cardiovascular risk.

EVALUATION STRATEGY AND METRICS

Evaluation will adopt both technical performance measures and interpretability assessments.

Confusion matrix:

Only when the prediction precisely anticipated in discretely of these four confusion matrix categories true positives, false negatives, true negatives and false positives and proportionately to the size of the dataset's positive and negative items separately, does it obtain an inflated score. The Matthews correlation coefficient (MCC) is the title of the statistical metric in examination.

Accuracy and F1 score calculated on confusion matrices are two of the almost widely used metrics for binary categorizing tasks. These statistical estimates, however, have a dangerous tendency to produce over-optimistic results, especially to use with unbalanced datasets. The G-mean is estimated to use the sensitivity and specificity function, according to.

$$G - mean = \sqrt{Sensitivity \times Specificity}$$

$$= \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \quad \dots\dots\dots (5)$$

$$Matthews \text{ Correlation Coefficient (MCC)}$$

$$= \frac{((TPTN) - (FPFN))}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad \dots\dots\dots (6)$$

Table 3.2 gives the short description of all evaluation metrics defined above through expressions (1) to (6).

Table 3 Evaluation metrics

Metric	Explanation
Precision	It decides how blameless the classifier is in identifying deceitful cases.
Recall	The superiority of a qualifier will be estimated by it.
Accuracy	It deals with the proficiency of the algorithm.
F-measure	It meets the necessities of the superiority of a classifier for the intermittent classes.
G-mean (Geometric mean)	It evaluates the performance of a classifier to create a balance between the minority and majority classes.
ROC (Receiver Operating Characteristics) Curve	It is used for computing the trade-offs among true positive and false positive fault rates in the occurrence of classification algorithms.
AUC (Area Under Curve)	It denotes the area that exists in a ROC curve.
ROC Convex Hull	It is used as a technique of detecting potentially optimal classifiers.

Models that were trained on highly unbalanced datasets can be assessed to use the balanced metric MCC. It measures the amount with which binary classifications predicted and seen correlate. The MCC's range is from -1 to +1. Whenever the coefficient is +1, the model is perfect, when it becomes 0 or same as a random hypothesis, once it is -1, the model is completely failed. The MCC metric is further reliable than the F1 score.

The Area in the Precision-Recall Curve (AUC-PR), which is the only value that reviews the precision recall (PR) curve, is another helpful indicator of the success of prediction when the classes are unbalanced. Instead of using the widely used receiver operating characteristic (ROC) curve, which could represent an excessively optimistic picture of the performance for an uneven dataset, PR curves are advised for highly skewed data. Then, using a binary classification problem, this score can be used to compare numerous models, with a score of 1.0 representing a model with perfect ability. Additionally, the area under the curve (AUC), which is often used in numerous other articles, is also measured for comparison purposes.

FUTURE SCOPE

Highly requirement of artificial intelligence and NLP for healthcare applications, indicates that we need to develop an explainable and clinically acceptable AI framework [22], which is specifically designed for cardiovascular text analytics. Formally research in this field is basically cauterized on structured data like ECG signals, echocardiographic images, laboratory parameters and utilizing deep learning architectures inclusive of CNNs and

RNNs [6][7][14]. We have to understand that these applications have very appropriate predictive performance but also limited transparency which reduces clinical confidence, interpretability and regulatory compliance. So upcoming research has to be focused on introducing an integrated XAI-NLP framework which is capable of understanding the insights within unstructured clinical text such as discharge summaries, physician notations and diagnostic reports. The models like BioBERT [8] and ClinicalBERT [9] which are Transformer-based biomedical dialect models, have shown an attractive performance in domain-specific attribute extraction but they have the issue with explanation of the decision making and works as the black box techniques. To overcome this issue we need to integrate NLP with ongoing explainability techniques such as LIME [10], SHAP [11], and Grad-CAM [12] with domain specific logical mechanisms. With the available structured medical knowledge from ontologies like UMLS and SNOMED CT, the interpretability layer reforms semantic alignment with cardiological reasoning. Frameworks in future need to be integrated with quantitative explainability metrics and qualitative human-centered validation by cardiologists to get explanations of results predicted by models that are clinically worthwhile and pertinent in real-world settings.

CONCLUSION

This review inspects closely and thoroughly that major difficulties and requirements with AI and NLP technologies have been rehabilitated in prediction and analysis of cardiovascular disorder, but clinically accepted and

10.48047/jocaaa.2024.33.08.475
transparent models remnants to be addressed. In today's systems, models preferred the accuracy of the prediction over the transparency of the result, especially when dealing with unstructured data. Due to which their use in real world clinical sceneries is affected by the deficiency of domain-integrated interpretability frameworks. To moderate the gap between model intelligence and clinical decision making pattern, explainable NLP models that also are integrated with ontology records and clinical validation are essential to be developed. This integration will expand model accuracy and also rise the clinical trust and acceptance for implementation in cardiovascular healthcare systems.

REFERENCES

- [1] World Health Organization, "Cardiovascular diseases (CVDs) fact sheet," 2024.
- [2] Naidoo V, Madamshetty L, Krishna SBN. Artificial intelligence for cardiology: from diagnosis to management. *Therapeutic Advances in Cardiovascular Disease*. 2026;20. doi:[10.1177/17539447251406847](https://doi.org/10.1177/17539447251406847)
- [3] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv preprint arXiv:1702.08608, 2017.
- [4] C. Madsen, M. R. Jørgensen, and O. Winther, "Post-hoc explanations fail to achieve their primary goal," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 932–941, 2022.
- [5] K. Sokol and P. A. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," in *Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, 2020, pp. 56–67.
- [6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 5998–6008.
- [7] P. Wagner, N. Strodthoff, R. Bousseljot, D. Kreiseler, F. Lunze, W. Samek, and T. Schaeffter, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 1, p. 154, 2020.
- [8] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.
- [9] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proc. Clin. NLP Workshop (ClinicalNLP)*, 2019, pp. 72–78.
- [10] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144.
- [11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 4765–4774.
- [12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.
- [13] M. N. S. Jahromi, H. B. Mofakhar, and A. H. K. Kazerouni, "SIDU-TXT: An explainable artificial intelligence algorithm for NLP models," *NLP Journal*, vol. 8, no. 2, pp. 101–114, 2024.
- [14] H. T. Nguyen, T. T. Nguyen, D. P. Tran, and M. N. Nguyen, "Explainable deep learning for ECG-based heart disease detection," *IEEE Access*, vol. 11, pp. 34312–34324, 2023.
- [15] Y. Kim, H. Park, J. Lee, and S. Kim, "Explainable AI-based clinical decision support systems: A systematic review," *Applied Sciences*, vol. 14, no. 2, p. 678, 2024.
- [16] A. E. W. Johnson, T. J. Pollard, L. Shen, H. L. Li-wei, M. Feng, M. Ghassemi, B. Moody et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.
- [17] P. Wagner, N. Strodthoff, R. Bousseljot, D. Kreiseler, F. Lunze, W. Samek, and T. Schaeffter, "PTB-XL dataset: A large publicly available ECG dataset," *Scientific Data*, vol. 7, p. 154, 2020.
- [18] R. Dey, S. Gupta, and A. K. Das, "AI-based prediction of heart failure using electronic health record data," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 4, pp. 1123–1135, 2023.
- [19] A. Paranjape, J. T. Zhou, and D. H. Park, "Information bottleneck for deep interpretability," in *Proc. Int. Conf. Machine Learning (ICML)*, 2020, pp. 7665–7675.
- [20] Rawat, S. S. and Mishra, A. K., "Review of Methods for Handling Class-Imbalanced in Classification Problems," arXiv preprint, arXiv:2211.05456, Nov. 2022.
- [21] Bisna, N. D., and Ajay James. "A privacy-preserved method for cardiovascular disease prediction from wearable device data." In *Advances in Electrical and Computer Technologies*, pp. 350–357. CRC Press, 2025.
- [22] S. Alghamdi, R. Mehmood, F. A. Alqurashi and A. Alzahrani, "Paving the Roadmap for XAI and IML in Healthcare: Data-Driven Discoveries and the FIXAIH Framework," in *IEEE Access*, vol. 13, pp. 174393–174427, 2025.