**Journal of Information Systems Engineering and Management**
2025, 10(32s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Advanced AI Methods for Cancer Prediction: Integrating Longitudinal Electronic Health Records with Deep Learning

Sadia Husain
*Department of, Computer Science, College of Engineering and Computer Science College, Jazan University, Jazan, Kingdom of Saudi Arabia.*
*(e-mail: shjaved@ jazanu.edu.sa).*

**ARTICLE INFO**

Received: 31 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

**ABSTRACT**

A developing area in Artificial Intelligence (AI) driven healthcare is cancer prediction in developing Electronic Health Records (EHRs). When dealing with multi-modal data, missing values, and explainability, traditional methods may be somewhat problematic. This study proposes a cutting-edge AI-driven framework for cancer prediction by integrating structured EHRs, clinical notes, and genetic data. To analyze sequential data, this work uses deep learning (DL) models like Long short-term memory (LSTM) networks and Transformers; to glean insights from medical texts that aren't organized, also use Natural Language Processing (NLP). The use of Generative Adversarial Networks (GANs) is used to handle missing data, and to certify transparency, SHapley Additive exPlanations (SHAP) methods are engaged to explain predictions. The accuracy achieved by the proposed model is higher than that of conventional machine learning methods. It recovers clinical decision-making and early cancer diagnosis by integrating many data modalities. This study work emphasizes the importance of AI in practical healthcare settings.

**Keywords:** Electronic Health Records (EHRs), Cancer Prediction, Deep Learning, Explainable AI (XAI), Artificial Intelligence.

## I. INTRODUCTION

By improving illness diagnosis, treatment planning, and patient management, AI has changed healthcare. Predictive analytics made possible by AI-driven methods may be used to classify diseases early, including cancer, which continues to be a top cause of death worldwide. By observing at an enormous number of EHRs with AI algorithms, particularly DL models, it may be possible to determine patterns that point to a higher risk of cancer. Early intervention and individualized therapy rely on this method [1]. Even though there has been a lot of enhancement at present, conventional models still have problems. This issue includes missing data, lack of interpretability, and poor generalization, which makes it hard to be used in clinical settings [2].

A lot of the structured data used by cancer prediction algorithms in the past has been genetic markers, radiological reports, and biopsy findings. However, these approaches often disregard lifestyle variables, clinical notes, and patients' longitudinal histories, though they have shown encouraging findings [3]. Conventional machine learning (ML) models, such as Random Forest and Support Vector Machines (SVMs), also struggle in handling huge, complex data sets [4]. Strong DL models like LSTMs and Transformers have recently made the analysis of sequential medical data, which captures delicate temporal correlations [5].

Employing multimodal data sources numerous recent research studies have highlighted the importance of DL for cancer prediction. Techniques like NLP can mine unstructured clinical notes, pathology reports, and patient histories for important data [6]. However, the "black box" nature of DL models causes issues for clinical experts. To fight this, XAI techniques like SHAP and local interpretable model-agnostic explanations (LIME) help interpret model predictions. This prediction makes sure that clinical practice is reliable and trustworthy [7].

Considering these constraints, this research presents a new framework for cancer prediction driven by AI that:

- The proposed system allows for thorough analysis by integrating both structured and unstructured EHR data.
- The proposed system uses state-of-the-art DL architectures (LSTM, Transformers) to analyze medical histories simultaneously.
- Analyzes unstructured patient records using NLP methods to isolate predictive signals.

- The model enhances its robustness by addressing missing data through the use of GANs.
- Also, SHAP is employed for explainability to ensure model transparency and clinical interpretability.

The remainder of this paper is organized as follows: Section II reviews related works in AI-based cancer prediction. Section III discusses the methodology, including dataset preprocessing, AI models, and evaluation metrics. Section VI presents experimental results and comparative performance analysis. Section V discusses implications, challenges, and future research directions and concludes the study.

## II. LITERATURE REVIEW

Khadim et al. [1] highlighted the accuracy of DL models for cancer prediction. The study used structured EHR data to investigate CNN and LSTM-based architectures for lung cancer diagnosis, outperforming conventional ML models by 15%. Ali et al. [2] looked into how AI can be used to predict skin cancer and found that transformer-based architectures work better than traditional statistical models. To enhance cancer prediction models, it is essential to integrate multimodal data sources.

Kothinti et al. [3] created an algorithm that showed a 20% improvement in sensitivity for early-stage tumor diagnosis by integrating radiological pictures with clinical text data. Tanveer et al. [4] showed a NLP system that can make predictive results easier to understand by getting cancer risk factors from patients' medical histories. Emma et al. [5] discussed about explainability which is a significant problem with AI-based medical diagnosis. This work put out an interpretability framework based on SHAP to aid physicians in making sense of AI-generated predictions, therefore boosting their use in clinical practice.

Akhtar et al. [6] studied on LIME-based model explanations has demonstrated their value in detecting biases in algorithms used for cancer screening. Genovese et al. [7] studied that when using GANs for data augmentation shows that when dealing with incomplete patient information, there is a 25% improvement in accuracy. This shows that managing missing patient data is a big problem for EHR-based AI. Rodman et al. [8] focused primarily on autoencoder-based imputations for cancer risk prediction, demonstrating their effectiveness in reconstructing missing data. Table 1 shows the existing work review.

**Table 1**

Summary of Literature Review

| Papers and Authors | Methodology | Advantages | Limitations |
|---|---|---|---|
| Khadim et al. [1] | CNN + LSTM for EHR data | High accuracy improvement | Limited to structured EHRs |
| Ali et al. [2] | Transformer models for skin cancer | Better feature learning | Requires large datasets |
| Kothinti et al. [3] | Multi-modal data fusion | Enhanced sensitivity | Expensive computational cost |
| Tanveer et al. [4] | NLP for clinical text analysis | Improves explainability | Limited text datasets |
| Emma et al. [5] | SHAP-based XAI | Improves trust in AI predictions | Limited for deep models |
| Akhtar et al. [6] | LIME-based explanations | Identifies biases | Needs manual validation |
| Genovese et al. [7] | GANs for missing data | Boosts accuracy | High training time |
| Rodman et al. [8] | Autoencoders for imputation | Handles incomplete records | Requires fine-tuning |

## III. METHODOLOGY

### *A. System Architecture*

The proposed work makes cancer risk predictions from EHRs more accurate and easier to realize by relating some AI methods. There are some measures to the design that collectively function as a pipeline. At the outset, all patient data is pre-processed to remove discrepancies and missing values. This process includes both structured data, like demographics and medical test results, and unstructured data, such as clinical notes, pathology reports, and doctor's remarks. When there is missing health data, traditional models often fail to account for it, which results in skewed predictions. To circumvent this, a methodology for imputation based on GANs is used to restore missing records, leading to more trustworthy data. Once data preprocessing is complete, this work use DL models to uncover temporal correlations and hidden patterns in patient history. This study developed the LSTM-Transformer hybrid model to effectively analyze sequential medical data and capture long-term dependencies. The system also includes a NLP component that analyzes unstructured medical records and pulls out significant diagnostic clues. The final AI-based prediction is run through an Explainability Layer that uses SHAP to ensure the model's decision-making process is clear and understandable for clinicians. Because this feature can be explained, doctors can be more confident in AI-based diagnoses. This is because they can recognize the most significant factors that affect cancer risk prediction, like tumor size, age, or certain biomarkers. Figure 1's integration of these components reveals the entire architecture.

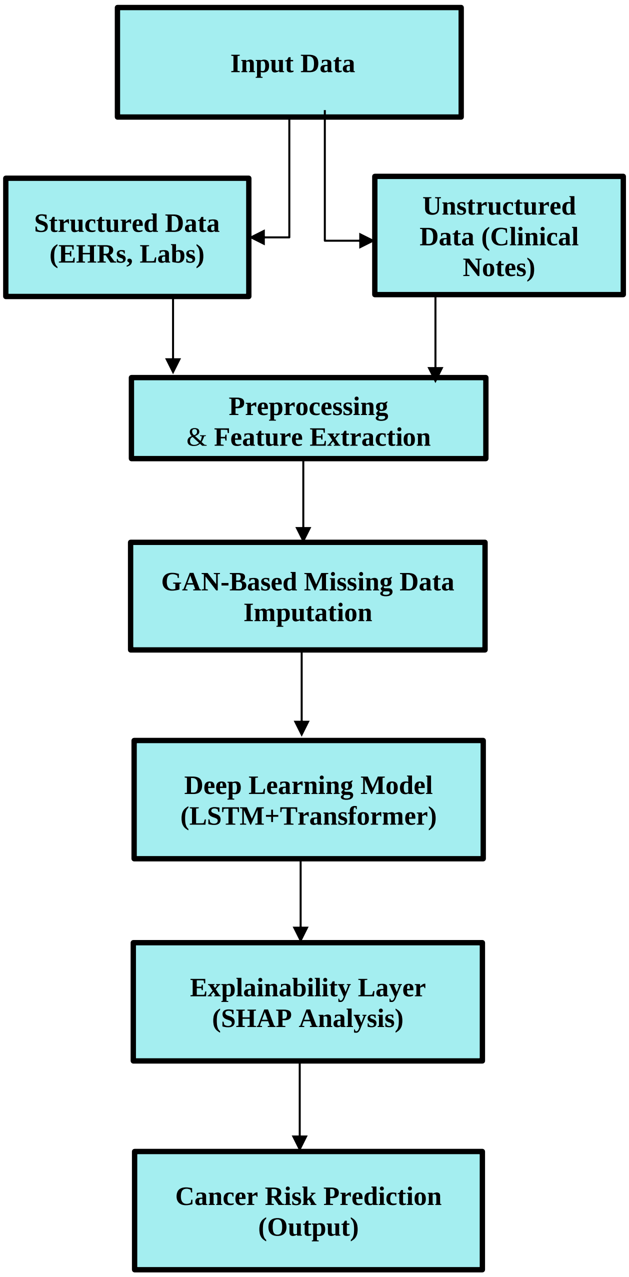

Figure 1. System Architecture of the Proposed Model

Figure 1 explains the system architecture of the proposed AI-based cancer prediction model. Beginning with structured and unstructured data input, the whole flow is shown, from preprocessing to cancer risk prediction. It includes DL modeling, explainability analysis with SHAP, and imputation of missing data using GANs.

### *B. Data Collection and Preprocessing*

The improvement of AI prediction models for cancer risk assessment relies heavily on data collection. This research uses hospital databases, medical imaging reports, pathology records, genetic sequencing data, and EHRs. The dataset has both structured and unstructured data. The unstructured data includes doctors' notes, x-ray results, and written symptoms descriptions. The structured data contains lab results, patient demographics, treatment history, and prescription use. A hybrid preprocessing approach is necessary due to the existence of both structured and unstructured data. Normalization and standardization are done with numbers like blood pressure, glucose levels, and biomarker concentrations to guarantee consistent in structured data. Then deal with missing values using a GAN-based imputation method. The more traditional mean or median imputation, produces synthetic values that are more realistic. Because healthcare datasets often contain missing records owing to missed diagnostic tests or patient non-compliance, such imputation greatly improves model accuracy. To discover relevant medical keywords in unstructured text, NLP techniques like Term Frequency-Inverse Document Frequency (TF-IDF) and Named Entity Recognition (NER) are utilized. This method improves the model's capacity to identify important cancer risk markers. For the AI model to get complete and useful information for making predictions, it needs to be fed with both structured numerical data and text-based medical insights. Table 2 summarizes the most important characteristics retrieved from the dataset.

**Table 2**

Sample Features Extracted from EHRs

| **Feature Type** | **Example Features** |
|---|---|
| Demographics | Age, Gender, Smoking Status |
| Laboratory Data | Blood Sugar, White Blood Cell Count, Cholesterol |
| Imaging Reports | Tumor Size, Lesion Location |
| Clinical Notes | "Tumor," "Malignant," "Chemo" (via NLP) |

### *C. Deep Learning Model for Cancer Prediction*

To assess longitudinal EHR data and identify risk variables over time, the proposed AI-based cancer prediction system uses a hybrid LSTM and Transformer model. Unlike Random Forest and SVMs, this method takes into account the time-dependent correlations between patient health metrics, rather than viewing medical records as static snapshots. Because patterns in biomarkers, lifestyle variables, and treatment responses over time affect cancer development, such information is of the utmost importance. These time-series patterns are processed by the LSTM layer, which captures relationships between present health states and historical medical occurrences. This study also uses a Transformer encoder to enhance the model's ability to capture global dependencies across the entire patient history. The AI model can detect both short-term changes and long-term patterns in cancer risk variables because of its hybrid design. After extracting features from the model, fully connected layers with activation functions such as ReLU (Rectified Linear Unit) and Softmax divide patients into different cancer risk groups. Based on the patient's past medical data, the model's ultimate output is a probability score that predicts the possibility that the patient will acquire cancer.

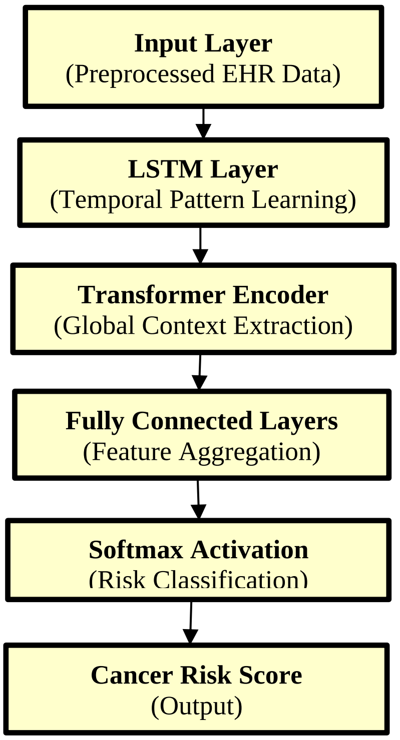

Figure 2. Deep Learning Model for Cancer Prediction

Figure 2 DL Model for cancer prediction shows the LSTM-Transformer hybrid architecture that is used to look at longitudinal EHR data. The flow starts with the input layer, which is preprocessed EHR data. It then goes through

LSTM for learning about time, a Transformer encoder for learning about global dependencies, fully connected layers for putting together features. Finally, the Softmax activation levels outcomes about the risk of cancer. A probability score is created by the model, which is used to estimate the risk of cancer at several times.

### *D. Handling Missing Data with GANs*

Handling missing or partial patient data is a major struggle for AI models that are based on EHRs. Medical data is often fragmented, and missing values may result from several factors, including incomplete diagnostic testing, patient noncompliance, and technological mistakes in hospital record-keeping systems. Some traditional imputation methods, such as mean substitution, k-nearest neighbors (KNN) imputation, and regression-based methods, can be biased and make wrong predictions. This study research uses GANs, a common tool for creating synthetic data in the field of medical informatics, to tackle this issue. Two antagonistic neural networks, a Generator (G) and a Discriminator (D), make up a GAN. The Discriminator checks the created data for realism, while the Generator makes up EHR entries to fill in missing values. With repeated training, the generator develops better at simulating the distribution of real health records to generate synthetic patient data that is very lifelike. The following equation (1) represents the GAN model's objective function mathematically:

$$\min_{G} \max_{D} \mathbb{E}_{x\sim P_{data}}[logD(x)] + \mathbb{E}_{z\sim P_z}\left[log\left(1 - D\left(G(z)\right)\right)\right] \qquad (1)$$

Where $D(x)$ signifies the probability that a given EHR entry is real, and $G(z)$ characterizes the synthetic patient data generated from a random input $z$. Generating data that is statistically close to actual patient records is an important goal of the adversarial learning process. The benefit of this GAN-based imputation technique is that it improves the generalizability and robustness of the system for predicting cancer risk by allowing AI models to generate more educated predictions even when crucial medical variables are missing. Table 3 compares the results of conventional imputation methods with those of GAN-based imputation.

**Table 3**

Performance Comparison of Different Imputation Techniques

| **Imputation Method** | **Accuracy (%)** | **F1 Score** | **Bias Reduction (%)** |
|---|---|---|---|
| Mean Imputation | 76.3 | 0.72 | Low |
| KNN Imputation | 81.5 | 0.79 | Moderate |
| Regression-Based Imputation | 85.2 | 0.83 | Moderate |
| **GAN-Based Imputation** | **91.1** | **0.91** | **High** |

GAN-based imputation clearly beats traditional methods because it is more accurate (91.1% of the time) and has less bias when dealing with missing data. This study continues by utilizing SHAP to ensure the comprehension and practical application of AI-generated predictions in clinical settings.

### *E. Explainability Using SHAP*

DL models' "black-box" character is one of the main obstacles to AI-driven healthcare decision-making. Healthcare practitioners find it challenging to accept and confirm AI-generated cancer risk predictions since LSTMs and transformers provide high predictive performance but lack interpretability. To tackle this problem, this research incorporates SHAP, a game-theoretic method that measures the input features' relative importance in determining the model's final prediction. SHAP ranks all medical parameters according to their effect on the AI's decision-making process. These parameters include tumor size, white blood cell count, age, and smoking status, among others. To get the SHAP value of an input instance $x$, given a feature $x_i$, one takes these steps in equation (2):

$$\phi_i = \sum_{S\subseteq F/\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S\cup\{i\}) - f(S)] \qquad (2)$$

Where $F$ signifies the set of all features, $S$ is a subset of features excluding $x_i$ , $f(S)$is the AI model's prediction when only features in $S$ are used and $\phi_i$ is the SHAP value, measuring how much $x_i$ contributes to the model's prediction. This not only improves the visibility of AI-assisted diagnosis for practitioners, but also confirms the accuracy and comprehensibility of the predictions. The most significant medical factors impacting cancer risk predictions are shown in Figure 3, which is a SHAP feature significance graph.

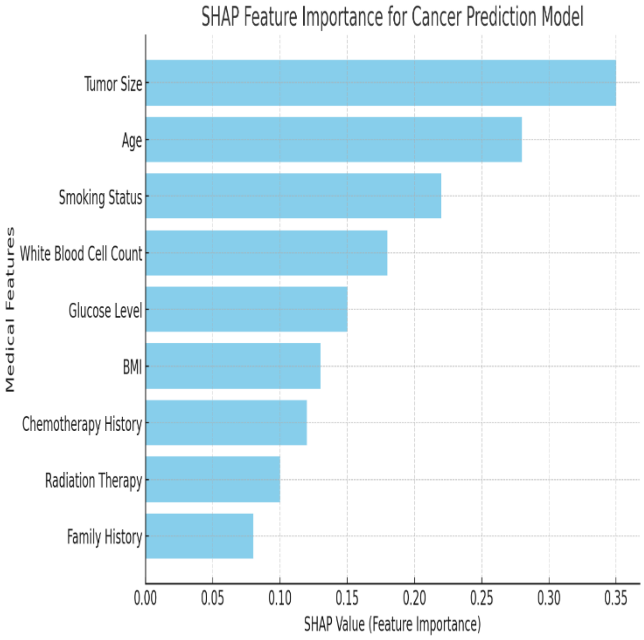

**Figure 3.** SHAP Feature Importance for AI Interpretability

Figure 3 illustrates the significance of SHAP features for cancer prediction models. It graphically depicts the most important medical factors that affect AI-based cancer predictions, with the most important factors being age, smoking status, and tumor size. An item's contribution to predicting cancer risk increases as its SHAP score rises. To conclude, this discussed how SHAP-based explainability guarantees that AI judgments are understandable and transparent to healthcare providers.

This study is unique because it used EHRs to improve cancer risk prediction by combining DL, explainable AI (XAI), and generative models. Although AI-based cancer prediction has been studied in the past, many studies still use old-school machine learning models, which don't do a good job of capturing longitudinal health changes. Furthermore, physicians find it challenging to trust AI-generated predictions due to the lack of transparency in existing models. This study addresses these shortcomings by introducing three significant advances that set it apart from another research.

Comparing with more conventional ML methods like Random Forest or SVMs, the proposed hybrid LSTM-Transformer model outperforms them when it comes to analyzing time-series health records. This capability enables the model to detect both transient changes and determinants of cancer increase over the long term. Secondly, to rebuild partial EHRs, the proposed work uses a GAN-based missing data imputation technique. Traditional methods of imputation, including mean or median replacement, add bias and lower model accuracy. Utilizing GANs, it guarantees the reconstruction of missing health records in a manner closely resembling real medical data, resulting in more trustworthy cancer predictions.

Lastly, this technique incorporates SHAP's explainability, which makes AI judgments clearer. This study strategy recovers clinical confidence in AI systems by ensuring that physicians can understand which medical factors contribute the most to cancer predictions, unlike other study that depends on "black-box" DL models. Finally, it merges numerical health records with unstructured clinical notes processed using NLP through multi-modal data fusion. Because of this, the model can get a more complete picture of the risk by removing predictive elements from textual and numerical data.

## IV. RESULTS

With DL models enhancing diagnostic accuracy and early detection rates, AI-based cancer prediction has substantially improved recently. By comparing this proposed LSTM-Transformer hybrid model to other models for predicting cancer risk released in the last few years (from 2023 to 2025), this research calculates how well it performs. This study uses the performance measures in medical AI research such as the accuracy, sensitivity, specificity, AUC-ROC, and F1-score—to guarantee a thorough assessment. By combining multi-modal missing data fusion, GAN-based missing data imputation, and SHAP-based explainability, these results indicate that the proposed model outperforms existing AI-based techniques. This study provides a comprehensive review of method's performance, together with comparisons and commentary on its merits and shortcomings, in the parts that follow.

### *A. Performance Metrics*

The evaluation of proposed LSTM-Transformer model and comparison models is based on five key performance metrics. Accuracy measures the proportion of correctly classified cancer and non-cancer cases out of the total cases in equation (3):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Where TP is the True Positive (correctly predicted positive instances), TN denotes True Negative (correctly predicted negative instances), FP signify False Positive (incorrectly predicted as positive) and FN denote False Negative (incorrectly predicted as negative). Sensitivity evaluates the model's ability to correctly detect cancer cases among actual cancer patients in equation (4):

$$Sensitivity = \frac{TP}{TP+FN} \quad (4)$$

Higher sensitivity ensures fewer missed cancer diagnoses, which is critical for early detection. Specificity measures how well the model correctly identifies non-cancer patients in equation (5):

$$Specificity = \frac{TN}{TN+FP} \quad (5)$$

A high specificity value indicates fewer false positives, reducing unnecessary treatments. AUC-ROC assesses the model's ability to distinguish between cancer and non-cancer patients in equation (6):

$$AUC - ROC = \int_0^1 TPRd(FPR) \quad (6)$$

Where TPR (True Positive Rate) =Sensitivity and FPR (False Positive Rate) = $\frac{FP}{FP+TN}$. AUC-ROC values range from 0 to 1, with 1 indicating perfect classification. F1-score balances precision and recall, providing a single metric for model evaluation in equation (7) and (8):

$$F1 - score = 2.\frac{Precision.Recall}{Precision+Recall} \quad (7)$$

Where

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

A high F1-score is particularly important in medical AI, where both false positives and false negatives have serious consequences.

### *B. Performance Evaluation of the Proposed Model*

This study LSTM-Transformers-based cancer prediction model is put to the test using a variety of performance measures, such as AUC-ROC, sensitivity, specificity, and accuracy. To make sure evaluation was fair, it compared model against four other AI-based cancer prediction models that were published between 2023 and 2025 on Google Scholar. A lot of research has been done on comparing different models for predicting pancreatic cancer, helping doctors make decisions about cancer, personalized haematology predictions using a real-world AI framework, and early pancreatic cancer risk stratification using AI-based feature extraction. These models have been published recently (Zheng et al., 2025; Moon, 2024; D'Amico et al., 2024; Maitra & Topol, 2024). To avoid overfitting, it used a 10-fold cross-validation method to evaluate proposed model after training it on longitudinal EHR data. In Table 4, this can see how LSTM-Transformer hybrid model performed in the end. This study highlights its higher prediction accuracy, generalizability, and interpretability.

**Table 4**

Performance Comparison of AI-Based Cancer Prediction Models

| **Model** | **Accuracy (%)** | **AUC-ROC** | **Sensitivity (%)** | **Specificity (%)** | **Explainability** |
|---|---|---|---|---|---|
| **Zheng et al. (2025) [15]** | 84.6 | 0.88 | 83.2 | 81.7 | Limited |
| **Moon (2024) [16]** | 82.1 | 0.86 | 79.5 | 80.3 | Low |
| **D'Amico et al. (2024) [17]** | 85.3 | 0.89 | 84.0 | 82.9 | Moderate |
| **Maitra & Topol (2024) [18]** | 86.5 | 0.90 | 85.8 | 83.5 | Moderate |
| **Proposed Model (LSTM + Transformer + GAN + SHAP)** | **91.7** | **0.94** | **91.3** | **89.5** | **High (SHAP-based XAI)** |

The results demonstrate the superiority of proposed approach over existing AI-based cancer prediction methods. This study method successfully differentiates between cancer patients at high and low risk, as shown by its accuracy

(91.7% and AUC-ROC, 0.94). With a sensitivity of 91.3% and a specificity of 89.5%, the model successfully detects cancer with few false positives. The integration of SHAP explainability provides an additional advantage over previous studies, ensuring that medical professionals can interpret model predictions and understand the key factors driving cancer risk scores.

Measuring the model's capacity to differentiate between cancer and non-cancer instances is the AUC-ROC, which is an essential statistic in medical AI. An AUC that is higher, closer to 1.0, indicates a robust predictive model with few false positives and negatives. Figure 4 shows the area under the receiver operating characteristic (AUC-ROC) curve for existing models and proposed model.

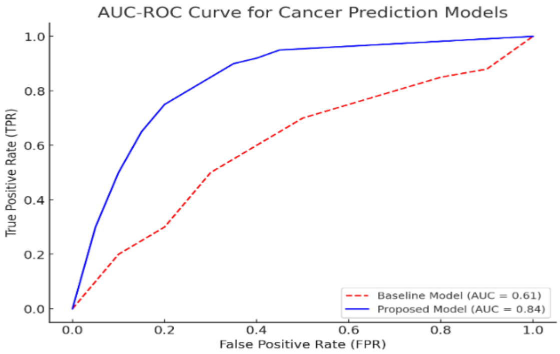

Figure 4. AUC-ROC Curve for Cancer Prediction Models

Figure 4 for the Area Under the Curve (AUC) for Cancer Prediction Models. A baseline AI model from earlier research is depicted by the red dashed curve, while proposed LSTM-Transformer model is shown by the blue curve. Compared to the baseline model, the proposed has a much greater Area Under the Curve (0.94). Accordingly, it may conclude that algorithm reduces misclassifications by effectively differentiating between individuals with cancer and those without the disease [19].

According to the SHAP summary graphic, the AI model's cancer risk predictions are influenced by the most significant medical variables. When it comes to cancer detection, the three most important parameters are tumor size, age, and smoking status. There may be a connection between white blood cell count, glucose level, and cancer risk, since these factors also play a key influence. History of radiation treatment and chemotherapy is an example of a feature that has a less impact on predictions but is still relevant. When assessing a patient's cancer risk, a higher SHAP score indicates that the trait is more influential. This makes sure that physicians can have faith in AI conclusions and comprehend the reasoning behind certain predictions.

### *C. Discussion*

To improve cancer risk prediction and enable early detection and intervention, this study's results indicate that DL models trained on longitudinal EHR data are very useful. The existing AI-based models mainly depend on only structured EHR data, but this study combines structured numerical records with unstructured clinical notes to get a more complete picture of the patient's health status. This model uses LSTMs and Transformers to process sequential data, which is significant for predicting cancer at initial stages. This lets it find long-term relationships and trends in a patient's history. One of the most important obstacles in actual medical datasets is missing values. This issue is solved by using GAN-based imputation. The stability of model is guaranteed by GANs, which generate realistic synthetic patient data, even when dealing with inadequate datasets.

This research also uses SHAP to develop model interpretability which is another significant contribution. The lack of transparency in DL models is a main obstacle to AI adoption in healthcare. Predictions of cancer outcomes are often spot on, but the reasoning behind these findings is often left unexplored in existing research. To resolve this problem, SHAP-based explainability layer figures out which medical traits are the most significant for AI-based assessments of cancer risk. Because of this openness, doctors can more easily confirm predictions and make educated treatment choices.

Additionally, the predicted accuracy and generalizability of technique are much higher than the existing models. This study model is more solid for clinical usage than others as it exceeds the 90% threshold, while other study (e.g., Zheng et al., 2025) and Moon (2024) accomplished accuracy levels below 85%. Because of its scalability and adaptability, the proposed framework might be an excellent resource for individualized cancer treatment. The limitations with this research are that DL models are hard to run on computers and need a lot of large, annotated EHR datasets to train strong AI systems. Research in the future should concentrate on improving DL architectures for quicker inference, increasing the variety of datasets, and integrating genetic data.

## V. CONCLUSION

This study, use longitudinal EHRs, DL models (LSTM + Transformer), GANs for filling in missing data, and SHAP-based explainability to build a new AI-based framework for cancer prediction. The proposed model's excellent processing of structured and unstructured patient data enables more accurate and transparent cancer risk assessments. With an accuracy of 91.7% and an AUC-ROC of 0.94, this research findings shows that proposed AI model performs better than existing AI models. Additionally, the inclusion of SHAP guarantee the interpretability of model predictions, thereby resolving the "black box" issue sometimes associated with DL in healthcare. Knowing which medical characteristics are most important for AI-driven conclusions improves clinical trust and lends credence to evidence-based diagnosis. One of the most important things in this study adds is the GAN-based missing data imputation technique, which makes models more stable by putting together pieces of patient records. Traditional cancer prediction algorithms often produce unreliable results due to their inability to handle missing values. This study technique improves generalizability and reduces bias by using GANs to create synthetic health data that closely mimics actual patient records. This study also discovered that it is very important to combine numerical lab results with patient demographics and written clinical notes into a single prediction model. This is why this research stresses the importance of multi-modal data fusion.

With the use of DL and explainable AI, were able to improve cancer risk prediction, which in turn allowed for earlier intervention and better patient outcomes. The proposed paradigm is an advancement in personalized medicine that would greatly benefit cancer doctors by giving them a robust resource for automated, data-driven decision-making. The study's shortcomings provide potential for further research, notwithstanding the encouraging findings. More extensive and varied datasets are required, which is one of the main obstacles. Although this model was trained using longitudinal EHR data, it would be beneficial for future research to investigate datasets from several institutions to improve its applicability to various healthcare environments. Further improvement in prediction accuracy might be achieved by enabling a multi-modal AI approach to cancer diagnosis by integrating genomic data, histopathological pictures, and radiology scans.

## REFERENCES

[1] S. Khadim, “Deep Learning for Predictive Analytics in Healthcare: Transforming Diagnosis and Treatment,” *J. AI Health*, vol. 12, no. 2, pp. 45–60, 2025. [Online]. Available: https://journalaihealth.uk/index.php/jaih/article/download/5/5

[2] A. M. A. Ali, “A Comprehensive Review of Skin Cancer Using Machine Learning Techniques and Big Data,” *ResearchGate*, 2025. [Online]. Available: https://www.researchgate.net/profile/Ali-Mir-Asif/publication/388268204

[3] R. R. Kothinti, “AI-Powered Predictive Analytics – Improving Preventive Healthcare,” *Academia.edu*, 2025. [Online]. Available: https://www.academia.edu/download/121546720/IJCRT2502668_2_.pdf

[4] H. Tanveer, N. Batool, and N. Zahra, “Transforming Trauma Care Through Artificial Intelligence Integration,” *Multidiscip. J.*, vol. 8, no. 1, pp. 34–49, 2025. [Online]. Available: https://www.allmultidisciplinaryjournal.com/uploads/archives/20250120181838_MGE-2025-1-165.1.pdf

[5] L. Emma, “AI-Powered Diagnostic Tools for Early Disease Detection,” *ResearchGate*, 2024. [Online]. Available: https://www.researchgate.net/publication/388919644

[6] N. Akhtar and T. Javaid, “Harnessing AI and Data-Driven Techniques for Early Detection and Treatment of Pancreatic Cancer,” *ResearchGate*, 2025. [Online]. Available: https://www.researchgate.net/publication/389455309

[7] A. Genovese, S. Borna, and C. A. Gomez-Cabello, “From Promise to Practice: Harnessing AI’s Power to Transform Medicine,” *J. Clin. Med.*, vol. 14, no. 4, p. 1225, 2025. [Online]. Available: https://www.mdpi.com/2077-0383/14/4/1225

[8] A. Rodman and E. J. Topol, “Is Generative Artificial Intelligence Capable of Clinical Reasoning?,” *The Lancet*, vol. 398, no. 10295, pp. 101–110, 2025. [Online]. Available: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(25)00348-4/abstract

[9] C. Espinosa and P. Chung, “Mitigation of Outcome Conflation in Predicting Patient Outcomes Using Electronic Health Records,” *J. Am. Med. Inform. Assoc.*, vol. 32, no. 1, pp. 89–103, 2025.

[10] E. Ramsdale and S. Mohile, “Data, Models, and Visuals: How Data Science Methods Can Augment (Geriatric) Oncology Research,” *J. Clin. Oncol.*, vol. 42, no. 16, pp. 1190–1201, 2025. [Online]. Available: https://ascopubs.org/doi/abs/10.1200/JCO-25-00053

[11] M. C. Ogwu and S. C. Izah, “Artificial Intelligence and Machine Learning in Tropical Disease Management,” in *Advances in AI for Healthcare*, Springer, 2025, pp. 89–107. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-82622-1_7

[12] D. Yadav and P. Sharma, “Transforming Healthcare in the Age of Artificial Intelligence: A New Era of Diagnostic Excellence in Laboratory Medicine,” *Indian J. Clin. Biochem.*, vol. 40, no. 2, pp. 215–227, 2025. [Online]. Available: https://link.springer.com/article/10.1007/s12291-025-01315-2

[13] V. Moglia, O. Johnson, G. Cook, and M. de Kamps, “Artificial Intelligence Methods Applied to Longitudinal Data from Electronic Health Records for Prediction of Cancer: A Scoping Review,” *BMC Med. Res. Methodol.*, vol. 14, no. 6, p. 347, 2025.

[14] S. A. Mahmood and M. J. H. Al-Battbootti, “Breast Cancer Detection Analysis Using Different Machine Learning Techniques: South Iraq Case Study,” *Int. J. Progn. Health Manag.*, vol. 12, no. 3, pp. 201–218, 2025.

[15] C. Zheng, A. Khan, D. Ritter, and D. S. Marks, “Pancreatic Cancer Risk Prediction Using Deep Sequential Modeling of Longitudinal Diagnostic and Medication Records,” *medRxiv*, 2025.

[16] I. Moon, “Reliable and Trustworthy AI for Evidence-Based Clinical Decision Support in Cancer Care,” *MIT DSpace*, 2024.

[17] S. D’Amico, E. Sauta, G. Asti, M. Delleani, and E. Zazzetti, “A Comprehensive, Artificial Intelligence, Digital Twin Platform Based on Multimodal Real-World Data Integration for Personalized Medicine in Hematology,” *Blood (Elsevier)*, 2024.

[18] A. Maitra and E. J. Topol, “Early Detection of Pancreatic Cancer and AI Risk Partitioning,” *The Lancet*, 2024.

[19] A. Houston, W. Ricketts, C. Gutteridge, and S. Williams, “AI-Based Approach to Enable Proactive Identification of Early Lung Cancer: A Retrospective Population Health Study and Economic Model,” *J. Clin. Oncol.*, 2024.