RESEARCH-ARTICLE

# Genomics-Enhanced Cancer Risk Prediction for Personalized LLM-Driven Healthcare Recommender Systems

**KEZHI LU**, University of Technology Sydney, Sydney, NSW, Australia

**JIE LU**, University of Technology Sydney, Sydney, NSW, Australia

**HANSHI XU**, University of Technology Sydney, Sydney, NSW, Australia

**KAIRUI GUO**, University of Technology Sydney, Sydney, NSW, Australia

**QIAN ZHANG**, University of Technology Sydney, Sydney, NSW, Australia

**HUA LIN**

**View all**



**Published:** 10 September 2025
**Online AM:** 20 June 2025
**Accepted:** 25 May 2025
**Revised:** 16 May 2025
**Received:** 29 September 2024

# Genomics-Enhanced Cancer Risk Prediction for Personalized LLM-Driven Healthcare Recommender Systems

KEZHI LU, JIE LU, HANSHI XU, KAIRUI GUO, and QIAN ZHANG, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia
HUA LIN and MARK GROSSER, 23Strands Pty Ltd, Sydney, Australia
YI ZHANG and GUANGQUAN ZHANG, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia

Cancer risk prediction is a cornerstone of personalized medicine that offers opportunities for early detection and preventive interventions. However, the current models are designed to predict cancer risk face several challenges. First, most rely on traditional statistical methods, which struggle to capture the complexity of genetic, family medical history, and lifestyle factors. Hence, the accuracy of these models is limited. Additionally, the models neglect to integrate multidimensional data sources, particularly genetic information like single nucleotide polymorphisms (SNPs), which could enhance prediction accuracy. Third, while the system might effectively predict risk, it cannot translate those predictions into actionable healthcare recommendations to reduce cancer risk.

In this study, we address all three of these limitations. With a focus on six prevalent cancers—we extracted SNP data from the UK Biobank and designed a novel risk prediction model for cancer and personalized healthcare recommendations based upon the mixture of experts (MoE) paradigm and large language models (LLMs), respectively. Named MoE-HRS, experts based two router networks for separate processing by the Transformer and the convolutional neural network (CNN). Experiments on UK Biobank data show that our model outperforms state-of-the-art cancer risk prediction models. To bridge the gap between risk prediction and practical healthcare applications, we devised a healthcare recommender system powered by LLMs. This approach holds promise for enhancing early detection rates and promoting preventive healthcare management (relevant coding and data are available at https://github.com/bjtu-lucas-nlp/MoE-HRS).

CCS Concepts: • **Information systems → Recommender systems**;

Additional Key Words and Phrases: Healthcare Recommender Systems, LLMs-Driven Recommender Systems, Genetic Risk Prediction, Recommendation

Funding support for this article was provided by the Australian Research Council (ARC) under Discovery Grant (DP220102635), Australian Research Council (ARC) under Linkage Scheme (LP210100414).
Authors' Contact Information: Kezhi Lu, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; e-mail: kezhi.lu-1@student.uts.edu.au; Jie Lu (corresponding author), Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; e-mail: jie.lu@uts.edu.au; Hanshi Xu, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; e-mail: hanshi.xu@student.uts.edu.au; Kairui Guo, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; e-mail: kairui.guo@uts.edu.au; Qian Zhang, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; e-mail: qian.zhang-1@uts.edu.au; Hua Lin, 23Strands Pty Ltd, Sydney, Australia; e-mail: hua.lin@23strands.com; Mark Grosser, 23Strands Pty Ltd, Sydney, Australia; e-mail: mark.grosser@23strands.com; Yi Zhang, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; e-mail: yi.zhang@uts.edu.au; Guangquan Zhang, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; e-mail: guangquan.zhang@uts.edu.au.


ACM 1558-2868/2025/9-ART152
https://doi.org/10.1145/3745022

**ACM Reference format:**
Kezhi Lu, Jie Lu, Hanshi Xu, Kairui Guo, Qian Zhang, Hua Lin, Mark Grosser, Yi Zhang, and Guangquan Zhang. 2025. Genomics-Enhanced Cancer Risk Prediction for Personalized LLM-Driven Healthcare Recommender Systems. *ACM Trans. Inf. Syst.* 43, 6, Article 152 (September 2025), 30 pages.
https://doi.org/10.1145/3745022

## 1 Introduction

Cancer remains one of the leading causes of morbidity and mortality worldwide, with an estimated 19.3 million new cases and 10 million cancer deaths in 2020 alone [5]. Early detection and prevention strategies play crucial roles in reducing cancer-related deaths and improving patient outcomes. However, the effectiveness of these strategies often depends on accurate risk assessment [31] and personalized healthcare recommendations.

The ability to accurately predict an individual's risk of cancer, cancer risk prediction, is a cornerstone of personalized medicine, enabling early detection and targeted preventive interventions. Such predictions mean preventative interventions can be implemented before a high-risk individual even develops cancer. Additionally, measures can be taken to detect cancer early in these individuals. By predicting an individual's likelihood of developing cancer, healthcare providers can adopt more targeted screening measures and offer tailored recommendations to reduce the impact of recommendations to improve the patients' outcome [15]. However, despite the importance of predicting cancer risk, existing models face significant limitations that hinder their effectiveness in real-world clinical applications [18, 43].

Traditional cancer risk prediction models are typically based on statistical approaches that rely heavily on epidemiological and lifestyle factors [40, 41]. While these methods have provided valuable insights, they often fail to account for the complex and multifaceted interactions between genetic, environmental, and behavioral risk factors, which have been shown to hold significant predictive potential in several applications [14]. **Single nucleotide polymorphisms (SNPs)**, when single nucleotide variations occurring in the genome, are the most common type of genetic variation among individuals and can provide important clues about a person's susceptibility to certain cancers [22]. Unfortunately, many existing models do not fully leverage this genetic information, resulting in suboptimal prediction accuracy.

**Genome-wide association studies (GWAS)** [24] have significantly advanced efforts to integrate genetic data into cancer risk prediction models. GWAS have revolutionized understanding of the genetic basis of complex diseases, including cancer, by identifying numerous genetic variants associated with the risk of the disease [22, 60]. They have uncovered thousands of SNPs that, while individually may have small effects, collectively contribute significantly to someone being susceptible to cancer. In addition, another key component of genetic-based cancer risk prediction is the **polygenic risk scores (PRS)** [7, 45], which aggregate the effects of multiple SNPs associated with cancer susceptibility into a single score that reflects an individual's genetic predisposition to a particular disease [22, 30, 61]. PRS models provide a quantitative measure of genetic predisposition, offering a more nuanced understanding of inherited cancer risk beyond traditional family history assessments. PRS have shown promise in stratifying individuals into different risk categories for various cancers, potentially enabling more targeted screening and prevention strategies [39]. However, despite their potential, the clinical application of PRS face several challenges:

— The complexity of genetic interactions: PRS typically assume that genetic variants have an additive effect, which may not capture the complex, non-linear interactions between genes and between genes and environmental factors [10].

—Population bias: Most large-scale GWAS have been conducted in populations of European ancestry, limiting the generalizability of PRS to diverse populations [38].
—Integration with other risk factors: While PRS provide valuable genetic risk information, they need to be integrated with other established risk factors (e.g., age, family history, lifestyle factors) so as to produce a comprehensive risk assessment [27, 29, 63].
—Interpretability and actionability: Translating a PRS into clinically actionable recommendations is difficult, as the relationship between genetic risk and specific preventive measures is not always straightforward [56].

To address these challenges, we propose MoE-HRS, a novel cancer risk prediction method that integrates genetic and other risk factors to enhance predictive accuracy while improving model interpretability and providing personalized healthcare recommendations. To effectively translate cancer risk assessments into actionable recommendations, healthcare recommender systems have emerged as an essential technology. These systems integrate risk prediction results with personalized medical guidance, suggesting tailored lifestyle modifications, screening schedules, and preventive treatments. However, most cancer risk prediction models operate in isolation, providing only a risk score without translating that score into actionable healthcare recommendations. This disconnect between risk prediction and practical healthcare guidance limits the clinical utility of such models. Patients and healthcare providers need more than just a risk estimate; they require personalized recommendations that inform screening, lifestyle adjustments, and preventive actions based on the predicted risk. The absence of such an integrated system remains a significant gap in current cancer risk prediction research.

In this study, we conducted experiments with six common cancer types that collectively account for a significant proportion of the global cancer burden: breast cancer, colorectal cancer, lung cancer, melanoma, non-Hodgkin's lymphoma, and prostate cancer. Our study aims to address these limitations by leveraging advanced deep-learning techniques to capture complex genetic interactions and integrate them with other risk factors. The detailed workflow of this work is shown in Figure 1.

In addition, the **mixture of experts (MoE)** paradigm [21] has emerged as a powerful technique for significantly scaling model capacity while maintaining minimal computational overhead. MoE has gained considerable attention in both academia and industry due to its ability to dynamically allocate model resources by activating only a subset of specialized "expert" networks for each input. This selective routing mechanism allows for vast model capacity, enabling better performance in handling complex tasks without proportionally increasing computational cost. By leveraging MoE techniques, models can balance scalability, efficiency, and performance, making this a key approach for large-scale machine learning applications [32, 51], particularly in domains such as **natural language processing (NLP)**, computer vision, and multi-task learning [13]. Each "expert" in the MoE model specializes in a different part of the input space, while a gating mechanism dynamically selects which experts are activated for a given input. This allows the model to maintain high expressiveness while keeping computation and memory use manageable. Inspired by this, we propose a novel methodology that combines advanced machine learning techniques for risk prediction with NLP for personalized healthcare recommendations. The resulting framework addresses the limitations of existing approaches. By designing an MoE framework using Transformers [59] and **convolutional neural network (CNN)** [28], our approach has the potential to capture complex non-linear relationships in genetic data that traditional PRS methods might overlook. Main architectures of MoE-HRS are:

—A cancer risk prediction model: Based on a novel MoE framework, this model fuses several Transformers with several CNN to leverage the strengths of both architectures:

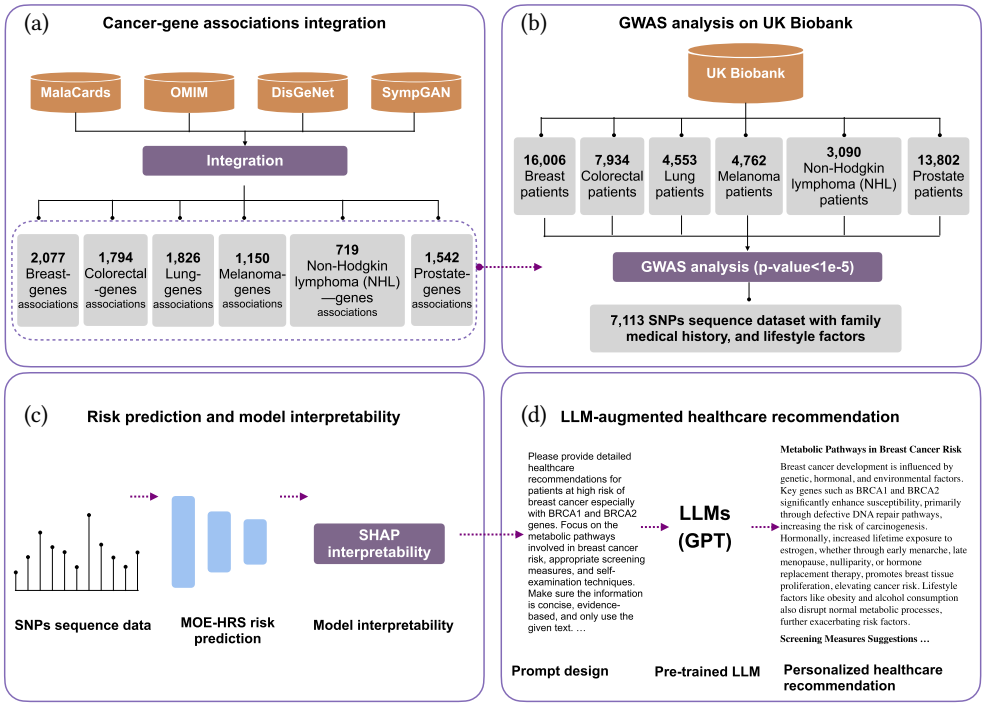


Fig. 1. The overview and workflow of entire work. (a) Data integration for cancer-gene associations. (b) GWAS analysis based on UK Biobank. (c) Cancer risk prediction based on MoE-HRS and model interpretability. (d) Large language model-driven healthcare recommendation.

  - The MoE Transformers excel at processing sequential data and capturing long-range dependencies, which is crucial for analyzing patient histories and temporal risk factors. These models are also particularly adept at handling complex genetic data by assigning specialized "experts" to different regions of the input space across different cancers. Thus, the model is free to adaptively focus on relevant features.
  - CNN are good at identifying spatial patterns, making them suitable for analyzing medical imaging data or other structured input features. Hence, CNN are adopted to analyze the spatial relationships between SNPs, which can reveal any underlying genetic structures associated with cancer susceptibility.
- Enhancing interpretability for cancer risk prediction: leverages **shapley additive explanations (SHAP)** to enhance interpretability by identifying key genetic and non-genetic factors contributing to cancer risk predictions, which is the key information in prompt designing for following **large language model (LLM)**-driven healthcare recommender systems.
- LLM-driven healthcare recommender systems: The LLMs generate personalized healthcare recommendations based on the risk prediction results. LLMs have demonstrated remarkable capabilities at understanding context and generating human-like text. This makes them ideal for providing tailored advice based on individual risk profiles.

As illustrated in Figure 1(a), we integrated data on cancer-gene associations from MalaCards [46],[1] OMIM [2],[2] DisGeNet [44],[3] and SympGAN [33].[4] In addition, as illustrated in Figure 1(b), we used SNP data from the UK Biobank [54],[5] to train and validate our model—the UK Biobank being a rich resource of genetic and phenotypic information. As illustrated in Figure 1(c), by leveraging this comprehensive dataset, we aimed to capture the complex interactions between genetic markers and cancer risk. Consequently, our model provides more nuanced and accurate predictions than traditional models.

The transformer-based MoE are particularly adept at handling complex genetic data. They work by assigning specialized "experts" to different regions of the input space, which then allows the model to adaptively focus on relevant features. Meanwhile, the CNN-based MoE analyze the spatial relationships between the SNPs to reveal the underlying genetic structures associated with cancer susceptibility. Through the design of these two MoE, MoE-HRS overcomes the limitations of existing approaches, providing a higher degree of predictive accuracy and robustness.

Beyond risk prediction, we extend the utility of our model by integrating it with a healthcare recommender system powered by an LLMs, as shown in Figure 1(d). The LLMs process the risk scores generated by our model and produces personalized healthcare recommendations tailored to the individual's risk profile. These recommendations cover preventive strategies, such as lifestyle modifications and dietary changes, as well as personalized screening schedules aimed at early detection. By offering actionable insights, our system bridges the gap between risk estimation and practical healthcare, making it a valuable tool for both patients and healthcare providers.

Including an LLM in the framework means the system can generate recommendations that are contextually relevant. Overall, the framework draws on vast medical knowledge to not only offer precise, data-driven advice, but also to ensure that individuals at higher risk receive tailored healthcare guidance that aligns with their unique needs. The result is a system that enhances patient engagement and potentially improves health outcomes. The integration of risk prediction and personalized healthcare recommendations represents a significant advancement in cancer prevention, addressing the limitations of current systems and fostering more proactive healthcare practices.

In summary, our study introduces a novel approach to cancer risk prediction and healthcare guidance. One of the key advancements is to use SNP data from the UK Biobank, family history records, and lifestyle factors combined with MoE Transformers and CNN for more accurate and reliable predictions. Adding an LLM-driven recommender system ensures that the predicted risk scores translate into meaningful and personalized healthcare advice, empowering patients to take control of their health and helping healthcare providers to make informed decisions. Our study aims to contribute to the ongoing efforts to improve cancer prevention, early detection, and positive patient outcomes. Our AI-driven framework ensures that individuals at higher risk receive tailored healthcare guidance that aligns with their unique needs, thus enhancing patient engagement and potentially improving health outcomes. The main contributions of this article are as follows:

—Design of a novel hybrid model for cancer risk prediction: To the best of our knowledge, we are the first to design a novel method that fuses MoE-based Transformers and CNN within a framework specifically designed for cancer risk prediction. The framework's architecture is tailored to capture both the contextual dependencies and spatial relationships within SNP

[1] https://www.malacards.org/.
[2] https://www.omim.org/.
[3] https://www.disgenet.com/.
[4] https://www.sympgan.org/.
[5] https://www.ukbiobank.ac.uk/.

data, leading to marked improvements in prediction accuracy across six major types of cancer: breast cancer, colorectal cancer, lung cancer, melanoma, non-Hodgkin's lymphoma, and prostate cancer.
— Advanced use of SNP data for more accurate predictions: The proposed model capitalizes on SNP data from the UK Biobank, which incorporates complex genetic variations to refine cancer risk prediction. By leveraging this large-scale genetic dataset, the model delivers more accurate and personalized risk assessments than traditional approaches.
— Development of an LLM-augmented personalized healthcare recommender system: Our LLM-augmented healthcare recommender system generates individualized healthcare guidance based on genetic risk profiles. This system provides actionable recommendations, such as preventive strategies, lifestyle adjustments, and tailored screening protocols, to bridge the gap between genetic risk assessment and clinical decision-making.
— A comprehensive comparison against other baseline models: We comprehensively evaluated MoE-HRS against a range of baseline models, including PRS-like, **support vector machine (SVM)**, **multi-layer perceptron (MLP)**, CNN, **long short-term memory (LSTM)**, **bidirectional LSTM (Bi-LSTM)**, **gated recurrent unit (GRU)**, Transformer, and SNP2Vec models. The proposed model generally outperforms these baselines in terms of key metrics such as **receiver operating characteristic area under the curve (ROC-AUC)**, precision, recall, and F1 score, demonstrating its superior predictive performance.

The remainder of this article is organized as follows. Section 2 reviews research on PRS, cancer risk prediction models, model interpretation, and healthcare recommender systems. Section 3 introduces the preliminaries of this article. Section 4 describes MoE-HRS in detail. Section 5 presents the dataset and experiments, where we compare MoE-HRS to existing approaches. An ablation study and hyper-parameters analysis are also included in this section. Finally, Section 6 concludes the article, the limitations of the study, and potential future directions of research.

## 2 Related Work

This section provides an overview of the related work in cancer risk prediction and personalized healthcare recommendations, focusing on four key areas: PRS, cancer risk prediction models, causality-aware analysis in model interpretation, and healthcare recommender systems.

### 2.1 PRS

PRS [45] have emerged as an efficient method for quantifying one's genetic susceptibility to complex diseases, including various types of cancer. These scores aggregate the effects of multiple genetic variants, primarily SNPs, to produce a number that reflects an individual's inherited risk of contracting a specific disease [56].

PRS have shown particular promise in identifying individuals at higher risk for common cancers. For example, Mavaddat et al. [39] demonstrated the efficacy of PRS in predicting breast cancer by combining the scores with known risk factors. As a result, Maddavat and colleagues were able to better stratify cancer risk in individuals. Klein et al. [23] showed that PRS could identify men at substantially increased risk of prostate cancer, potentially informing screening strategies. In addition, Jia et al. [20] developed a PRS that significantly improved risk predictions when combined with a family history and lifestyle factors. Kachuri et al. [22] demonstrated that integrating PRS with modifiable risk factors improves cancer risk prediction. These studies consistently show that incorporating PRS into cancer risk models significantly improves predictive accuracy, especially when combined with traditional risk factors like family history, age, and lifestyle factors.

However, despite their growing popularity and demonstrated utility, PRS-based models face several challenges:

—Gene–Gene and Gene–Environment Interactions: PRS typically assume that genetic variants have an additive effect, which may not capture complex interactions between genes or between genes and environmental factors [62].
—Population Bias: Most PRS models have been developed based on populations of European ancestry, limiting their generalizability to more diverse populations [38]. If not addressed, this bias can lead to inaccurate predictions.
—Integration with Other Risk Factors: While PRS provide valuable genetic risk information, they must be effectively integrated with other established risk factors for comprehensive risk assessment [56].
—Interpretability and Clinical Actionability: Translating PRS into clinically actionable recommendations remains challenging, as the relationship between genetic risk and specific preventive measures is not always straightforward [29].

To address these shortcomings, machine learning-based approaches, such as those proposed in this study, offer a promising solution by better capturing the complex interactions within genetic data and across diverse population groups. In contrast to the static nature of PRS, our model leverages MoE Transformers and CNN, which dynamically adapt to various genetic and contextual data, providing a more nuanced and accurate risk assessment.

### 2.2 Cancer Risk Prediction Models

Cancer risk prediction is a rapidly evolving field. The aim is to identify individuals at high risk of developing cancer by leveraging genetic, clinical, and environmental data.

Traditional risk prediction models have relied heavily on epidemiological factors, such as family history, age, smoking status, and exposure to environmental carcinogens. For example, Gail et al. [11] draw on factors such as age, family history, and reproductive history to predict the risk of contracting breast cancer. Tammemägi et al. [55] designed the PLCO Model for predicting lung cancer, which analyzes smoking history, age, race, education, BMI, family history, and the presence of chronic lung disease. However, while these models have proven useful in some circumstances, they often fail to capture the full complexity of cancer risk. This is because they do not draw on genetic data and, moreover, they cannot model non-linear relationships between risk factors.

In recent years, the advent of machine learning, and deep learning has opened new avenues for improving cancer risk prediction and personalized healthcare [1, 57]. Such techniques have offered improvements in predictive power because they are generally driven by large, multidimensional datasets. Kourou et al. [26] reviewed various machine learning methods designed for cancer prediction and prognosis, finding their performance to be superior to that of traditional models. In addition, deep learning models, particularly CNN, have been successfully employed to analyze medical images and genetic data for stratifying cancer risk [17].

The use of genomic data has signaled another significant advancement in the field. In particular, the use of SNP data has drastically improved predictions over an individual's cancer susceptibility, especially those models that also incorporate PRS [7]. On the downside, these models often struggle with the complexity of the data. Hence, we need better ways of handling multidimensional data.

The MoE-HRS approach introduced in this study, which integrates MoE Transformers with CNN, overcomes these challenges by effectively learning from both genetic and contextual data to provide a more accurate and personalized prediction of cancer risk. Our model also improves upon existing methods by leveraging SNP data from the UK Biobank, a large, diverse dataset, to further enhance its predictive capabilities.

### 2.3 Causality-Aware Analysis in Model Interpretation

As machine learning models become increasingly complex and ubiquitous in decision-making processes, the need to be able to interpret the results these models produce has grown significantly. Model interpretation, also known as **explainable AI (XAI)** [64], refers to the techniques and methods that help humans understand and trust the predictions made by machine learning models. This is especially important in high-stakes applications like healthcare, where the decisions made by models directly affect patient outcomes, or in finance, where model predictions can influence credit risk assessments and regulatory compliance.

Traditional machine learning models, such as linear regression and decision trees, are inherently interpretable due to their simple, transparent structures. However, modern models like deep neural networks, ensemble methods, and Transformers, which excel at handling large-scale, high-dimensional data, are often considered to be "black boxes." The opaqueness of these models poses challenges in terms of transparency, trust, fairness, and accountability. As a result, the field of model interpretation has gained significant attention, with various techniques developed to explain the inner workings and predictions of complex machine learning models. For example, LIME [48] provides local explanations by approximating the predictions of a complex model with an interpretable linear model in the vicinity of a specific instance. LIME generates perturbations around the instance, evaluating the black-box model on these perturbations, and fits a simple model to explain the prediction locally. More recently, the SHAP method[6] [36] has become one of the most popular methods for explaining the predictions of machine learning models.

Recent advances in XAI have increasingly focused on moving beyond correlational explanations toward causal understanding. Pearl's causal hierarchy [4] provides a framework for distinguishing between associational, interventional, and counterfactual reasoning, with the latter two being essential for true causal understanding. In the context of model interpretation, several studies have attempted to bridge this gap between correlation and causation.

Janzing et al. [19] proposed a causal view of feature importance, arguing that SHAP values can be interpreted as a form of counterfactual analysis that quantifies how model predictions would change if features were altered while holding others constant. This perspective aligns SHAP with fundamental causal inference concepts, making it valuable for exploring potential causal relationships captured by machine learning models.

In the medical domain, Lundberg et al. [37] applied SHAP to interpret complex models for predicting patient outcomes, demonstrating how feature importance can suggest potential causal mechanisms in healthcare applications. Similarly, other researchers have explored how model interpretation can inform causal understanding in cancer prediction [68].

More recently, LLMs have emerged as powerful tools for enhancing model interpretability by generating natural language explanations of model predictions [66]. These explanations can incorporate domain knowledge about causal mechanisms, potentially bridging the gap between statistical patterns and causal understanding. Our work builds upon these advances by integrating SHAP-based causal analysis with LLM-generated explanations to provide causally informed interpretations of cancer risk predictions.

### 2.4 Healthcare Recommender Systems

Healthcare recommendation systems aim to provide individuals and healthcare providers with actionable insights based on medical data, risk factors, and predicted health outcomes. These systems are increasingly being powered by machine learning and deep learning models that analyze large volumes of data to offer personalized healthcare advice.

[6] http://github.com/slundberg/shap.

Traditional healthcare recommendation systems typically rely on rule-based algorithms or decision trees, offering generalized advice based on predefined health guidelines. While these systems provide useful information, they often lack personalization, leading to recommendations that may not fully align with the unique needs of individual patients. Take, for example, clinical decision support systems that use if-then rules to guide clinical decisions [52]or evidence-based medicine guidelines that provide standardized recommendations based on population-level data [50].

While these systems provide useful information, they often lack personalization, leading to recommendations that may not fully align with the unique needs of individual patients. With the advancement of deep learning and NLP, AI-powered healthcare recommender systems have become more sophisticated [34, 35]. LLMs, such as the GPT-based models, are now capable of generating personalized recommendations by analyzing a patient's medical history, lifestyle, and genetic information. For instance, Singhal et al. [53] demonstrated the potential of LLMs to provide accurate medical advice based on clinical data, highlighting their applicability in healthcare. Meanwhile, Yao et al. [65] designed an ontology-aware prescription recommender system based on multi-evidence healthcare data.

In the context of cancer prevention and management, healthcare recommender systems can play a pivotal role in providing patients with tailored advice regarding lifestyle modifications, screening schedules, and preventive measures:

—Lifestyle modifications include personalized recommendations for diet, exercise, and other lifestyle factors based on individual risk profiles.
—Screening schedules refers to tailored screening recommendations that take personal risk factors and genetic predisposition into account.
—Preventive measures means specific advice on preventive actions or medications based on individual risk assessments.

However, many current systems fail to incorporate individual cancer risk predictions into their recommendations, limiting their effectiveness in providing truly personalized advice. Our study addresses this gap by integrating cancer risk prediction with LLM-driven healthcare recommendations. By using the risk scores generated by our MoE Transformer and CNN model, the LLMs are prompted to deliver precise, personalized healthcare guidance based on an individual's specific cancer risk profile. This integrated approach ensures that patients receive actionable, data-driven advice aligned with their unique genetic and clinical characteristics. This represents a significant advancement over existing healthcare recommendation systems.

In summary, while significant progress has been made in the fields of PRS, cancer risk prediction models, and healthcare recommender systems, each area faces challenges in handling the complexity and personalization required for optimal cancer care. Our proposed model, which integrates MoE Transformers, CNN, and LLMs, builds on this foundational work and addresses these limitations by offering a comprehensive, personalized solution for both accurate cancer risk prediction and personalized healthcare recommendations. This novel approach holds the potential to enhance early detection efforts, improve patient outcomes, and revolutionize preventive healthcare strategies.

## 3 Preliminaries

In this section, we introduce the notations and terminology used throughout the article, followed by a detailed description of the research target. Table 1 lists the main notations of this article used.

The goal of this research is to improve risk prediction by leveraging the MoE framework, which is known for its efficiency in handling large-scale and complex models. In traditional risk prediction models, fixed feature sets and static model structures are often unable to capture the complex,

Table 1. Notations

| Notation | Definition |
|---|---|
| $\mathbf{X} \in \mathbb{R}^{L \times d}$ | Input genetic sequence of length $L$ and dimension $d$ |
| $G_i$ | The gating network of the $i$th expert |
| $G_i^{tran}$ | The gating network of the $i$th transformer-based expert |
| $G_i^{cnn}$ | The gating network of the $i$th CNN-based expert |
| $G_i(\mathbf{X})$ | The output of the gating function for the $i$th expert |
| $E_i$ | The $i$th expert |
| $E_i^{tran}$ | The $i$th transformer-based expert |
| $E_i^{cnn}$ | The $i$th CNN-based expert |
| $E_i(\mathbf{X})$ | The output of the $i$th expert |
| $\mathbf{E}$ | The input embedding of Transformer-based MoE |
| $Embed_{pos}(p)$ | The position embeddings of the tokens in the input sequence |
| $Embed(\mathbf{X})$ | The embeddings of each token in the input sequence |
| $\mathbf{W}$ | The trainable parameter matrics |
| $\mathbf{b}$ | The bias vector |
| $\hat{r}$ | The final risk prediction |

non-linear relationships present in high-dimensional genetic, family medical history, or lifestyle data. To address these limitations, we propose using MoE to dynamically allocate specialized sub-networks (experts) for different data distributions or subsets of the feature space, allowing for more nuanced modeling of diverse risk factors.

The primary research target is to enhance risk prediction accuracy by:

Efficiently scaling model capacity: MoE allows the model to increase capacity without a linear increase in computation, as only a small subset of experts is activated for each input. This selective approach enables the model to handle complex data distributions while maintaining computational efficiency dynamically. Capturing non-linear relationships: By dynamically routing inputs to specialized experts, the model can better capture non-linear relationships that may be missed by traditional linear or static models, particularly in cases where certain risk factors interact differently across different population subgroups. Improving interpretability: Through the gating mechanism, it is possible to gain insights into which subsets of the model (experts) are activated for specific types of risk predictions, thereby enhancing the interpretability of the model's predictions, which is critical in domains like healthcare.

In summary, this research aims to be the first to explore the application of MoE in the domain of risk prediction, pushing the boundaries of traditional modeling techniques by integrating a dynamic, scalable, and interpretable approach. Through comprehensive experimentation and analysis, we seek to demonstrate the effectiveness of MoE in delivering more accurate and reliable risk predictions across various domains.

## 4 Methodology

In this study, we present a new approach, MoE-HRS (Figure 2 and Algorithm 1), to genetic risk prediction by designing a MoE method based on transformer architecture and CNN. Our method represents the first attempt to utilize the MoE paradigm on cancer risk prediction, potentially offering improved accuracy and interpretability in genomic analysis.

### 4.1 Overview of the Proposed Method

As illustrated in Figure 2 and Algorithm 1, our proposed method leverages the strengths of both transformer models and CNN, combined within an MoE framework. This approach allows for the

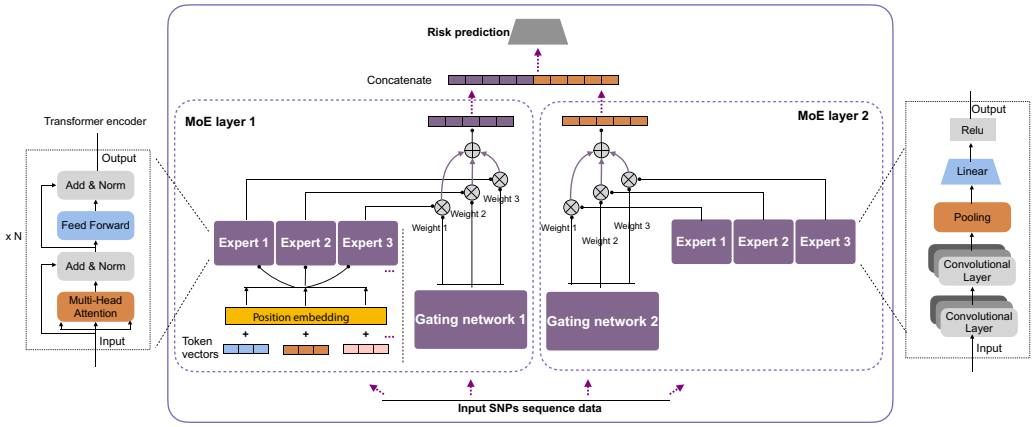


Fig. 2. The proposed model architecture of MoE-HRS. Overall, it includes two MoE layers (MoE layer 1 and MoE layer 2) and two gating networks. The MoE layer 1 is mainly based on transformer encoder. The MoE layer 2 is mainly based on CNN, pooling layer, and linear layer.

capture of complex, long-range dependencies in genetic data while also benefiting from the local feature extraction capabilities of CNN. The MoE architecture enables the model to dynamically route input data to specialized "expert" sub-networks, potentially improving performance on heterogeneous genetic datasets. Let $\mathbf{X} \in \mathbb{R}^{L \times d}$ represent an input genetic sequence of length $L$ and dimension $d$. Our MoE model can be formulated as:

$$f(\mathbf{X}) = \sum_{i=1}^{N} G_i(\mathbf{X}) \cdot E_i(\mathbf{X}),$$

where $N$ is the number of experts, $G_i(\mathbf{X})$ is the gating function for the $i$th expert, and $E_i(\mathbf{X})$ is the output of the $i$th expert. This formulation allows the model to dynamically combine the outputs of multiple specialized expert networks, each potentially focusing on different aspects of the genetic data. The gating function determines the relevance of each expert for a given input, enabling the model to adapt to various genetic patterns and structures. As shown in Figure 2, our MoE-HRS mainly consists of two MoE layers—transformer-based MoE networks and CNN-based MoE network.

## 4.2 Transformer-Based MoE Networks

The transformer-based MoE networks are designed to capture long-range dependencies and complex interactions within genetic sequences. Each transformer-based expert consists of the following components.

*4.2.1 Input Embedding:* $\mathbf{E} = Embed(\mathbf{X}) \in \mathbb{R}^{L \times d_{\text{model}}} + Embed_{pos}(p) \in \mathbb{R}^{L \times d_{\text{model}}}$. This step converts the input genetic sequence into a dense vector representation. Each element of the sequence is mapped to a high-dimensional vector $Embed$ and $Embed_{pos}$, allowing the model to learn rich representations of genetic elements. $Embed_{pos}$ represents the position embeddings of the tokens in the input sequence.

**Algorithm 1:** MoE-HRS: Genomics-Enhanced Cancer Risk Prediction with SHAP Analysis and LLM-Driven Healthcare Recommender Systems

**Require:** Sequential input: SNPs data $X_{SNP}$, lifestyle data $X_{life}$, family history $X_{fam}$, labels $Y$
**Ensure:** Trained MOE model, SHAP-based SNPs/Gene importance feature analysis, and LLM-generated healthcare recommendation

**Initialization:** Define $T$ Transformer layers, $C$ CNN layers, and $M$ experts
**Input Processing:** Encode $X_{SNP}$, $X_{life}$, and $X_{fam}$
**for** each training epoch **do**
    **Transformer Processing:**
    **for** $t = 1$ to $T$ **do**
        $E_t^{tran} = \text{TransformerLayer}(E_{t-1}^{tran})$
    **end for**
    **CNN Feature Extraction:**
    **for** $c = 1$ to $C$ **do**
        $E_c^{cnn} = \text{CNNLayer}(E_{c-1}^{cnn})$
    **end for**
    **Expert Selection:** Compute gating network output
    **for** $m = 1$ to $M$ **do**
        $\alpha_m = \text{softmax}(G_m(E_T^{tran}, E_C^{cnn}))$
    **end for**
    **Expert Predictions:** Compute weighted sum
    $\hat{r} = \sum_{m=1}^{M} \alpha_m E_m(E_T^{tran}, E_C^{cnn})$
    **Loss Calculation:** Compute loss $\mathcal{L}\text{risk}(\theta)$
    **Backward Propagation:** Update model parameters
**end for**
**Trained Model:** Save trained MOE Transformer-CNN model
**Cancer Risk Prediction:**
**for** each patient $x$ **do**
    Compute risk score $\hat{r} = \text{MoE-HRS}(x)$
    Output patient-specific cancer risk prediction
**end for**
**SHAP Analysis:**
**for** each patient $x$ **do**
    Compute SHAP values for input SNPs and genes
    Identify top contributing SNPs and genes related to cancer risk
**end for**
**Prompt Design for LLM-based Healthcare Recommendation:**
**for** each patient $x$ **do**
    Extract top contributing SNPs and genes from SHAP analysis
    Construct prompt with risk level, significant SNPs, and lifestyle factors
    Query LLM with the generated prompt to obtain personalized healthcare recommendations
    Output LLM-generated healthcare advice
**end for**
**Return** trained model, SHAP importance analysis results, patient risk predictions, and LLM-generated recommendations

*4.2.2 Transformer Encoder Layers.* For each layer $l$:

$$\begin{aligned}\mathbf{A}_l &= \text{MultiHeadAttention}(\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l)\\ \mathbf{B}_l &= \text{LayerNorm}(\mathbf{A}_l + \text{Residual}_l)\\ \mathbf{F}_l &= \text{FFN}(\mathbf{B}_l)\\ \mathbf{H}_l &= \text{LayerNorm}(\mathbf{F}_l + \mathbf{B}_l),\end{aligned}$$

where $\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l$ are query, key, and value matrices derived from the previous layer's output. These layers form the core of the transformer architecture. The multi-head attention mechanism allows the model to focus on different positions and capture various types of dependencies. The layer normalization and residual connections help in training deep networks, while the feed-forward network adds non-linearity to the model.

*4.2.3 Multi-Head Attention.*

$$\text{MultiHeadAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\mathbf{W}^O,$$

where $h$ denotes the total number of multi-heads, $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are query, key and value matrices, $\mathbf{W}^O$ denotes the trainable parameters, each head is computed as:

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V).$$

Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. This is particularly useful for capturing various types of dependencies in genetic sequences. The detailed attention function follows:

$$Attention(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) = softmax(\mathbf{Q}\mathbf{W}_i^Q \mathbf{K}\mathbf{W}_i^{K^T}/\sqrt{d})\mathbf{V}\mathbf{W}_i^V.$$

*4.2.4 Position-Wise Feed-Forward Network.*

$$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2,$$

where $\mathbf{W}_1, \mathbf{W}_2$ are trainable parameters, $\mathbf{b}_1, \mathbf{b}_2$ denote the bias matrics. This network applies the same feed-forward operation to each position separately and identically. It introduces non-linearity and allows the model to transform the representations at each position.

### 4.3 CNN-Based MoE Networks

The CNN-based MoE are designed to identify local patterns and structural features within genetic sequences. As shown in Figure 2, they comprise:

*4.3.1 Convolutional Layers.* For each layer $l$:

$$\mathbf{C}l = \text{ReLU}(\text{Conv1D}(\mathbf{X}l - 1, \mathbf{W}_l) + \mathbf{b}_l).$$

These layers apply convolution operations to extract local features from the input sequence. The ReLU activation introduces non-linearity, allowing the network to learn complex patterns.

*4.3.2 Pooling Layers.*

$$\mathbf{P}_l = \text{MaxPool}(\mathbf{C}_l).$$

Pooling layers downsample the feature maps, retaining the most salient features and reducing computational complexity.

*4.3.3 Fully Connected Layers.* For the final layer:

$$\mathbf{F} = \text{ReLU}(\mathbf{P}_L \mathbf{W}_F + \mathbf{b}_F).$$

The fully connected layer integrates features from all positions, providing a global view of the extracted patterns.

### 4.4 Gating Network

The gating network is a crucial component of our MoE architecture, determining how to route input data to various experts. It is defined as:

$$g_i(\mathbf{X}) = \frac{\exp(h_i(\mathbf{X}))}{\sum_{j=1}^{N} \exp(h_j(\mathbf{X}))},$$

where $h_i(\mathbf{X})$ is the output of a small neural network for expert $i$. This softmax formulation ensures that the gating weights sum to 1, effectively creating a probability distribution over the experts. The gating network learns to assign higher weights to experts that are more relevant for a given input. The small neural network we used here for expert $h_i(\mathbf{X})$ takes an input $\mathbf{X}$ and produces a vector of gating coefficients g:

$$h_i(\mathbf{X}) = W_g \mathbf{X} + b_g,$$

where $W_g \in \mathbb{R}^{n \times d}$ represents the weight matrix of gating network of $n$ experts and dimension $d$, $b_g \in \mathbb{R}^{n \times d}$ denotes the bias vector.

### 4.5 Integration and Output Layer

The final output is computed by combining the expert outputs weighted by their gating values:

$$\hat{r} = \mathbf{W}o \cdot concat\left(\sum_{i=1}^{N} G_i^{tran}(\mathbf{X}) \cdot E_i^{tran}(\mathbf{X}), \sum_{i=1}^{N} G_i^{cnn}(\mathbf{X}) \cdot E_i^{cnn}(\mathbf{X})\right) + \mathbf{b}_o,$$

where *concat* is concatenation method to concatenate the transformer-based MoE output and the CNN-based MoE output, $G_i^{tran}, G_i^{cnn}$ represent the gating network of these two MoE, $E_i^{tran}, E_i^{cnn}$ represent the expert network of these two MoE, $\hat{r}$ denotes the final risk prediction. This formulation allows the model to adaptively combine the predictions of different experts based on their relevance to the input.

### 4.6 Training Procedure

We optimize the model parameters $\theta$ by minimizing the loss function:

$$\mathcal{L}(\theta) = \mathcal{L}\text{risk}(\theta),$$

where $\mathcal{L}$risk is the risk prediction loss. The risk prediction loss is defined as:

$$\mathcal{L}\text{risk}(\theta) = \frac{1}{n} \sum_{i=1}^{n} (r - \hat{r})^2,$$

where $\mathcal{L}\text{risk}(\theta)$ is the MSE loss, $n$ represents the total number of dataset.

### 4.7 Feature Importance Analysis and LLM-Enhanced Interpretability

*4.7.1 SHAP-Based Feature Importance Analysis.* We implemented SHAP analysis [36] to quantify feature importance in our cancer risk prediction model. This approach provides insights into how different genetic and clinical features contribute to the predicted risk scores, enhancing model transparency.

For each patient, we computed SHAP values across all input features, identifying which factors most strongly influence the risk predictions. By aggregating these values across patients, we determined which genetic markers consistently show high importance in the model's decision-making process. This analysis reveals the statistical relationships between specific features and cancer risk predictions, providing a foundation for more meaningful interpretation.

*4.7.2 LLM-Enhanced Explanations Establishing Causality-Aware Relationships.* Our methodology establishes causality-aware relationships between genetic/lifestyle factors and cancer risk through a novel integration of SHAP analysis with medical knowledge via LLMs. This approach bridges the gap between risk scores and causality-aware associations in two key ways:

First, our SHAP analysis identifies which genetic and lifestyle factors consistently demonstrate strong relationships with cancer risk predictions across diverse patients. When certain factors consistently show high importance across different patient contexts, this suggests robust relationships that may reflect underlying causal mechanisms rather than spurious correlations.

Second, we systematically connect these statistical patterns to established medical knowledge by integrating StatPearls information into our LLM prompts. StatPearls is a comprehensive, peer-reviewed medical knowledge resource that provides detailed information about disease mechanisms, including how specific genetic factors are associated with cancer development and progression. These entries often describe the biological relationships between genetic variants and cancer, including pathways and mechanisms that suggest potential causal connections. By incorporating references to this medical knowledge alongside SHAP results, our method contextualizes the statistical patterns identified by our model within established medical understanding documented in the literature.

Through this integration of SHAP analysis, medical knowledge, and natural language explanation, our methodology advances beyond mere correlation toward a more causally informed understanding of cancer risk factors. While acknowledging the limitations of observational data for establishing definitive causation, our approach represents a significant step toward translating statistical predictions into causally meaningful insights that can inform clinical decision-making and personalized cancer prevention strategies.

## 5 Experiments

### 5.1 Dataset and Preprocessing

To begin our experiments, we first collected cancer-gene associations. As shown in Figure 1(a), these associations were selected from MalaCards, OMIM, DisGeNet, and SympGAN. Combining the data from these four databases resulted in 2,077 breast cancer-gene associations, 1,794 colorectal-gene associations, 1,826 lung cancer-gene associations, 1,150 melanoma-gene associations, 719 non-Hodgkin's lymphoma-gene associations, and 1,542 prostate cancer-gene associations. As illustrated in Figure 3, we performed an overlap analysis on the gene association data for the six cancers. The complete dataset is available at https://github.com/bjtu-lucas-nlp/MoE-HRS. We then incorporated the SNP data from the UK Biobank based on these cancer-gene associations, as illustrated in Figure 1(b). These data contain a wide range of genetic and other information, including SNPs alongside family history (illnesses of mother, father, and siblings), and lifestyle factors (smoke, alcohol intake). Thus, our dataset was highly suitable for predicting the risk of certain cancers.

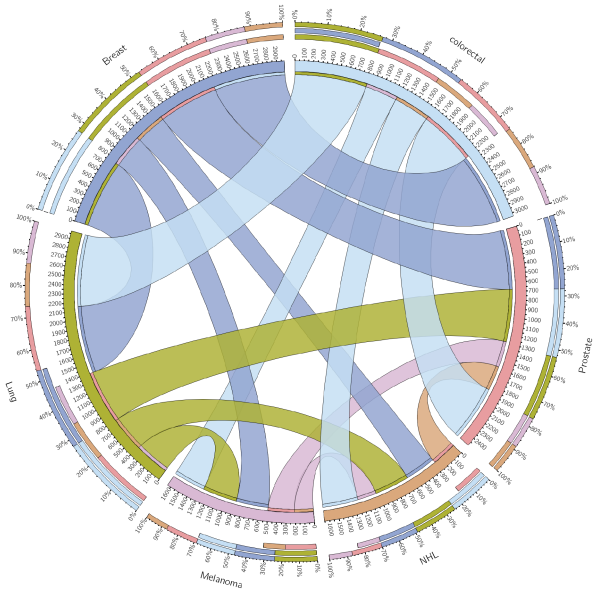


Fig. 3. Overlapped genes analysis based on six cancer types.

We specifically focused on the following six cancer types: breast cancer, colorectal cancer, lung cancer, melanoma, non-Hodgkin's lymphoma, and prostate cancer. The "UKB(iobank) imputation from genotype" data were used for further analysis. The cancer information was obtained from the "Cancer register" item. Only individuals with malignant cancers were considered as having cancer, while non-cancerous individuals were those who neither had malignant cancers nor benign tumors. The British ethnic background was kept for further analysis. For gender-specific cancers, like prostate cancer, only male samples were retained. For cancers with a significant gender imbalance, like breast cancer, only female samples were kept.

We used the plink2 software[7] to process the UK Biobank genotype data. First, all the SNPs with a reference frequency of between 0.4 and 0.6, and with a reference/alternative pair of A/T, T/A, C/G or G/C were considered ambiguous and removed. SNPs in the gene body and potential promoter region were included (i.e., 2,000 base pairs upstream of each gene location). These SNPs, as well as cancerous and non-cancerous samples, were extracted from the UK Biobank data. SNPs with a minor allele frequency greater than 0.001 were kept. Those individuals in whom more than 90% of SNPs were detected as well as those SNPs detected in more than 90% of individuals were kept. SNPs were filtered out according to the Hardy-Weiberg equilibrium ($1 \times 10^{-5}$). Here, logistic regression was applied. Finally, we filtered out data with p-values less than 1e-5 as a threshold. The genotype data was then converted into a pure data matrix.

*5.1.1 SNPs Selection.* We selected the SNPs most relevant to each cancer type based on existing GWAS and the previous literature. Using filters to remove low-quality SNPs, we retained around 7,113 SNPs across the six cancer types. This provided a robust set of genetic markers for risk prediction.

*5.1.2 Data Splitting.* The dataset was randomly split into training (70%), validation (15%), and test (15%) sets, ensuring a balanced distribution of cases and controls for each cancer type.

[7] https://www.cog-genomics.org/plink/2.0/.

Table 2. Statistics of the Datasets

| Dataset | # Breast | # Colorectal | # Lung | # Melanoma | # NHL | # Prostate |
|---|---|---|---|---|---|---|
| Patient | 16,006 | 7,934 | 4,553 | 4,762 | 3,090 | 13,802 |
| Non-patient | 16,361 | 7,911 | 4,627 | 4,764 | 3,121 | 13,772 |
| Total | 32,367 | 15,845 | 9,180 | 9,526 | 6,211 | 27,574 |

The SNP data was encoded into binary format (0, 1, 2 for genotype representation) and normalized for compatibility with training the deep learning model. In addition to training and evaluating the model, this dataset was also used for ablation experiments to assess the contribution of different input features and hyper-parameter analysis to fine-tune the model's performance. The detailed statistics of the dataset are shown in Table 2.

### 5.2 Baseline Models

To compare the performance of our hybrid model, we evaluated it against the following baseline models. The reason we selected these baselines is primarily due to the fact that many risk prediction models do not publicly share their specific code or data, making it challenging to choose suitable baselines. As a result, we opted for the following algorithms as baseline comparison models and have also made their code publicly available at https://github.com/bjtu-lucas-nlp/MoE-HRS.

(1) PRS-Like [45]: A linear model that aggregating the effects of multiple genetic variants into a single score that reflects an individual's genetic predisposition to a particular disease. In our experiment, we screened the SNP data through the cancer-gene relationship data and then conducted the PRS method experiment instead of conducting the experiment on the whole genome data, so we named it PRS-like.
(2) SVM [9]: A type of supervised learning algorithm mainly utilized for classification tasks. They operate by identifying the optimal hyperplane that effectively divides data points of different classes in a high-dimensional space. SVMs are especially adept at managing high-dimensional data and are recognized for their resilience against outliers.
(3) MLP [49]: A type of feedforward neural network consisting of multiple layers of interconnected nodes, where each node is a neuron that applies a nonlinear activation function. It is commonly used for supervised learning tasks such as classification and regression, learning complex patterns in data through backpropagation.
(4) CNN [28]: Predominantly used for image processing tasks but have also proven effective in various other domains, including text classification and time-series analysis. CNN utilize convolutional layers to automatically extract features from input data, allowing them to recognize patterns and structures efficiently.
(5) LSTM [16]: A specialized type of **recurrent neural network (RNN)**, that is effective at learning long-term dependencies in time series or sequential data. LSTMs address the vanishing gradient problem commonly encountered in traditional RNNs, enabling them to capture temporal patterns over extended sequences.
(6) Bi-LSTM [12]: The model takes the LSTM a step further that processes sequences in both forward and reverse directions, providing additional context. This bidirectional approach allows the model to consider context from both past and future time steps, improving its understanding of sequential data and leading to enhanced performance in tasks such as sentiment analysis and machine translation.
(7) GRU [8]: An RNN variant used to capture temporal dependencies in sequential data and time-series tasks.

(8) Transformer [59]: Revolutionized the field of NLP by introducing a novel architecture based on self-attention mechanisms. Unlike traditional sequential models, Transformers can process entire sequences simultaneously, enabling them to capture complex dependencies and relationships within the data. This architecture has led to significant advancements in tasks such as machine translation, text generation, and summarization.
(9) SNP2Vec [6]: A Scalable Self-Supervised Learning Approach for Risk Prediction. SNP2Vec utilizes self-supervised learning to transform high-dimensional SNP data into informative vector representations. By pre-training on large-scale genomic datasets, it captures the complex patterns and relationships among genetic variants, enhancing predictive modeling for disease risk assessment. This method enables efficient and scalable SNP representation learning, improving the accuracy of risk prediction models in genetics and precision medicine.

### 5.3 Evaluation Metrics

We used the following metrics to evaluate model performance:

—Precision: The ratio of true positives to the sum of true and false positives. High precision indicates that a model has a low false positive rate, making it particularly valuable in applications where the cost of false positives is high, such as in medical diagnoses or fraud detection.
—Recall: The ratio of true positives to the sum of true positives and false negatives. High recall is crucial in situations where missing a positive instance is significantly detrimental, such as in disease detection or security screening.
—F1 Score: The harmonic mean of precision and recall, providing a balanced measure of model performance. A higher F1 score indicates a better balance between precision and recall, making it an important metric in contexts where both false positives and false negatives carry significant consequences.
—ROC-AUC: A fundamental evaluation metric used to assess the performance of binary classification models. It provides a comprehensive measure of a model's ability to discriminate between positive and negative classes across all classification thresholds.

### 5.4 Computational Complexity Analysis

Our proposed MoE-HRS model integrates Transformer and CNN expert networks for cancer risk prediction. This section analyzes the computational complexity of the model architecture.

Assuming input data with dimension $d$ and sample size $n$, we analyze the computational complexity of each MoE-HRS component:

(1) *Gating Network*:
—Assuming the gating network is a simple feed-forward neural network, its complexity is $O(d \cdot k)$, where $k$ is the number of experts
(2) *Transformer Expert Network*:
—With $L$ Transformer layers, each having $h$ attention heads and model dimension $d_{model}$
—Self-attention mechanism complexity: $O(n^2 \cdot d_{model})$
—Feed-forward network complexity: $O(n \cdot d_{model}^2)$
—Overall Transformer expert complexity: $O(L \cdot (n^2 \cdot d_{model} + n \cdot d_{model}^2))$
(3) *CNN Expert Network*:
—With $L_c$ convolutional layers, each having $F_i$ filters of size $K \times K$
—Single convolutional layer complexity: $O(F_i \cdot F_{i-1} \cdot K^2 \cdot n_i)$, where $n_i$ is the feature map size at that layer
—Overall CNN expert complexity can be approximated as $O(\sum_{i=1}^{L_c} F_i \cdot F_{i-1} \cdot K^2 \cdot n_i)$

Table 3. Risk Prediction Results for Breast Cancer

| | Precision (± Std) | Recall (± Std) | F1 Score (± Std) | ROC-AUC (± Std) |
|---|---|---|---|---|
| PRS-like | 0.6316 (0.007) | 0.0008 (0.004) | 0.0015 (0.003) | 0.4870 (0.005) |
| SVM | 0.6496 (0.006) | 0.5992 (0.004) | 0.6234 (0.007) | 0.6968 (0.006) |
| MLP | 0.6264 (0.005) | 0.5520 (0.006) | 0.5868 (0.005) | 0.6601 (0.006) |
| CNN | **0.6737** (0.006) | 0.6503 (0.004) | 0.6618[a] (0.007) | 0.7371[a] (0.008) |
| LSTM | 0.5266 (0.005) | 0.5303 (0.007) | 0.5285 (0.006) | 0.5541 (0.006) |
| Bi-LSTM | 0.5380 (0.006) | 0.5759 (0.004) | 0.5563 (0.005) | 0.5729 (0.005) |
| GRU | 0.5450 (0.007) | 0.5132 (0.006) | 0.5286 (0.008) | 0.5765 (0.007) |
| Transformer | 0.6484 (0.006) | 0.6615[a] (0.005) | 0.6549 (0.05) | 0.7276 (0.007) |
| SNP2Vec | 0.6462 (0.004) | 0.6613 (0.006) | 0.6537 (0.005) | 0.7080 (0.005) |
| **MoE-HRS** | 0.6671[a] (0.004) | **0.6956** (0.005) | **0.6706** (0.004) | **0.7471** (0.004) |

Results are reported as Mean ± Std over five runs.
Values in bold represent the best result, and those with an asterisk ([a]) represent the second-best result.

(4) *Expert Fusion*:
—Expert fusion complexity: $O(k \cdot d_{output})$, where $d_{output}$ is the output dimension

The total computational complexity of MoE-HRS is the sum of these components. Compared to single Transformer or CNN models, MoE-HRS typically activates only a subset of expert networks during inference, which means that despite the increase in total parameters, the actual computational complexity does not grow linearly.

### 5.5 Overall Performance

In this section, we present the experimental results of the cancer risk prediction task, comparing the performance of the proposed MoE-HRS model with nine baseline models: SVM, MLP, CNN, LSTM, Bi-LSTM, GRU, Transformer, and SNP2Vec. The evaluation metrics include precision, recall, F1-score, and ROC-AUC, which provide a comprehensive analysis of each model's prediction ability.

Tables 3–8 present the detailed experimental results for all six cancer types. All models were trained on our final dataset comprising SNPs alongside age, family history, and lifestyle factors, which is described in Sections 5.1. We utilized an 70-15-15 training-validation-test split for model evaluation. To ensure the robustness of our results, we conduct each experiment five times with different random seeds. We report the mean and standard deviation (Mean ± Std) of the evaluation metrics to reflect the stability of our model. Overall, our MoE-HRS model outperforms all the baseline models across ROC-AUC, precision, recall, and F1 score, which demonstrates the model's ability to capture both contextual and spatial information in SNP data significantly improved cancer risk prediction.

Specifically, as shown in Table 3, in breast cancer, MoE-HRS leads the second-best result by about 0.01 to 0.034 in Recall, F1 Score, and ROC-AUC. Only the precision indicator lags behind the best performing CNN method by 0.004, ranking second best. Table 4 shows the risk prediction results of different algorithms for colorectal cancer. Our MoE-HRS algorithm outperforms other algorithms in all four indicators. It is worth noting that in the Recall and F1 score results, the MoE-HRS algorithm is 16% and 8.1% higher than the second best Transformer algorithm, respectively. In the lung cancer risk prediction experiment, the MoE-HRS method was ahead of the second-best CNN algorithm by approximately 0.7% to 4.4% in all four indicators. As shown in Table 6, in the risk prediction of melanoma cancer, the precision and recall of the MoE-HRS algorithm reached 0.8254

Table 4. Risk Prediction Results for Colorectal Cancer

| | Precision (± Std) | Recall (± Std) | F1 Score (± Std) | ROC-AUC (± Std) |
|---|---|---|---|---|
| PRS-like | 0.5027 (0.004) | 0.5771 (0.006) | 0.5373 (0.007) | 0.4974 (0.005) |
| SVM | 0.5598 (0.007) | 0.5353 (0.006) | 0.5473 (0.005) | 0.5763[a] (0.006) |
| MLP | 0.5462 (0.004) | 0.4766 (0.005) | 0.5090 (0.004) | 0.5571 (0.004) |
| CNN | 0.5606[a] (0.003) | 0.4884 (0.004) | 0.5220 (0.004) | 0.5734 (0.003) |
| LSTM | 0.5164 (0.004) | 0.5009 (0.006) | 0.5086 (0.005) | 0.5124 (0.005) |
| Bi-LSTM | 0.5250 (0.003) | 0.5253 (0.005) | 0.5251 (0.004) | 0.5207 (0.004) |
| GRU | 0.5170 (0.004) | 0.4981 (0.005) | 0.5074 (0.003) | 0.5220 (0.004) |
| Transformer | 0.5329 (0.005) | 0.6483[a] (0.006) | 0.5850[a] (0.005) | 0.5520 (0.005) |
| SNP2Vec | 0.5309 (0.004) | 0.6421 (0.005) | 0.5812 (0.004) | 0.5722 (0.004) |
| **MoE-HRS** | **0.5621** (0.004) | **0.7527** (0.005) | **0.6324** (0.003) | **0.6001** (0.004) |

Results are reported as Mean ± Std over five runs.
Values in bold represent the best result, and those with an asterisk ([a]) represent the second-best result.

Table 5. Risk Prediction Results for Lung Cancer

| | Precision (± Std) | Recall (± Std) | F1 Score (± Std) | ROC-AUC (± Std) |
|---|---|---|---|---|
| PRS-like | 0.5140 (0.007) | 0.1768 (0.006) | 0.2631 (0.005) | 0.4963 (0.005) |
| SVM | 0.7663 (0.009) | 0.7445 (0.008) | 0.7553 (0.007) | 0.8305 (0.007) |
| MLP | 0.8653 (0.005) | 0.8133 (0.006) | 0.8385 (0.004) | 0.9034 (0.005) |
| CNN | 0.9548[a] (0.007) | 0.9465[a] (0.004) | 0.9507[a] (0.006) | 0.9734[a] (0.005) |
| LSTM | 0.5677 (0.008) | 0.5906 (0.006) | 0.5789 (0.005) | 0.5912 (0.005) |
| Bi-LSTM | 0.5977 (0.006) | 0.5644 (0.004) | 0.5806 (0.004) | 0.6222 (0.005) |
| GRU | 0.5491 (0.007) | 0.5804 (0.005) | 0.5643 (0.005) | 0.6174 (0.005) |
| Transformer | 0.5335 (0.006) | 0.5393 (0.004) | 0.5364 (0.005) | 0.5584 (0.004) |
| SNP2Vec | 0.7079 (0.005) | 0.8439 (0.006) | 0.7699 (0.005) | 0.8349 (0.006) |
| **MoE-HRS** | **0.9966** (0.004) | **0.9683** (0.005) | **0.9823** (0.005) | **0.9900** (0.004) |

Results are reported as Mean ± Std over five runs.
Values in bold represent the best result, and those with an asterisk ([a]) represent the second-best result.

and 0.7540 respectively. In the risk prediction experiment results of non-Hodgkin's lymphoma cancer, MoE-HRS was 13.08%, 0.72%, 14.95%, and 15.87% higher than the second-best model in Precision, Recall, F1 Score, and ROC-AUC indicators, respectively. Table 8 shows the risk prediction results of prostate cancer. The MoE-HRS algorithm is better than the second-best model in Recall, F1 score and ROC-AUC by 0.0646, 0.0278 and 0.0063, respectively.

In addition, we observed significant variations in the predicted outcomes for different cancers. For instance, MoE-HRS achieves accuracies of 0.9966, 0.8254, and 0.7270 for lung cancer, melanoma, and non-Hodgkin's lymphoma, respectively. However, for breast cancer, colorectal cancer, and prostate cancer, the highest accuracies are limited to 0.6737, 0.5621, and 0.6247, respectively. Similarly, MoE-HRS achieves high Recall scores for certain cancers: 0.7527 for colorectal, 0.9683 for lung, and 0.7540 for melanoma. In contrast, for breast cancer, non-Hodgkin's lymphoma, and prostate cancer, the highest Recall scores are 0.6956, 0.6615, and 0.6500. Likewise, for colorectal cancer, the F1 score and ROC-AUC results are relatively low than other types cancer, influenced by the precision results. Specifically, the highest F1 score achieved is 0.6324, and the best ROC-AUC result is 0.6001.

Table 6. Risk Prediction Results for Melanoma Cancer

| | Precision (± Std) | Recall (± Std) | F1 Score (± Std) | ROC-AUC (± Std) |
|---|---|---|---|---|
| PRS-like | 0.5266 (0.007) | 0.1829 (0.005) | 0.2715 (0.006) | 0.5075 (0.006) |
| SVM | 0.6593 (0.005) | 0.6642 (0.004) | 0.6618 (0.003) | 0.7237 (0.004) |
| MLP | 0.6522 (0.007) | 0.6336 (0.006) | 0.6427 (0.004) | 0.7089 (0.005) |
| CNN | 0.7335[a] (0.006) | 0.7297[a] (0.008) | 0.7316[a] (0.006) | 0.8119[a] (0.007) |
| LSTM | 0.5669 (0.005) | 0.5998 (0.006) | 0.5829 (0.005) | 0.6020 (0.005) |
| Bi-LSTM | 0.5772 (0.006) | 0.5766 (0.005) | 0.5769 (0.004) | 0.6073 (0.005) |
| GRU | 0.5552 (0.006) | 0.5696 (0.006) | 0.5623 (0.006) | 0.5685 (0.005) |
| Transformer | 0.5761 (0.005) | 0.5037 (0.006) | 0.5375 (0.005) | 0.5905 (0.005) |
| SNP2Vec | 0.6459 (0.004) | 0.6800 (0.005) | 0.6626 (0.004) | 0.7252 (0.006) |
| **MoE-HRS** | **0.8254** (0.004) | **0.7540** (0.003) | **0.7881** (0.004) | **0.8586** (0.004) |

Results are reported as Mean ± Std over five runs.
Values in bold represent the best result, and those with an asterisk ([a]) represent the second-best result.

Table 7. Risk Prediction Results for Non-Hodgkin's Lymphoma Cancer

| | Precision (± Std) | Recall (± Std) | F1 Score (± Std) | ROC-AUC (± Std) |
|---|---|---|---|---|
| PRS-like | 0.6092 (0.006) | 0.0172 (0.006) | 0.0334 (0.007) | 0.4772 (0.006) |
| SVM | 0.5883 (0.004) | 0.5326 (0.005) | 0.5591 (0.004) | 0.5925 (0.003) |
| MLP | 0.6077 (0.007) | 0.5512 (0.005) | 0.5873 (0.005) | 0.6199 (0.006) |
| CNN | 0.6429[a] (0.005) | 0.5450 (0.009) | 0.5899 (0.005) | 0.6371 (0.007) |
| LSTM | 0.5566 (0.006) | 0.6568[a] (0.007) | 0.6026 (0.005) | 0.5675 (0.006) |
| Bi-LSTM | 0.5413 (0.005) | 0.6413 (0.006) | 0.5871 (0.006) | 0.5520 (0.005) |
| GRU | 0.5621 (0.004) | 0.5431 (0.006) | 0.5524 (0.005) | 0.5498 (0.006) |
| Transformer | 0.5301 (0.004) | 0.5062 (0.003) | 0.5179 (0.005) | 0.5298 (0.004) |
| SNP2Vec | 0.6107 (0.003) | 0.6382 (0.005) | 0.6241[a] (0.004) | 0.6409[a] (0.004) |
| **MoE-HRS** | **0.7270** (0.004) | **0.6615** (0.003) | **0.6927** (0.005) | **0.7382** (0.004) |

Results are reported as Mean ± Std over five runs.
Values in bold represent the best result, and those with an asterisk ([a]) represent the second-best result.

Additionally, we found that both the CNN model and MoE-HRS achieved accuracy and recall above 0.9465 in the lung cancer risk prediction experiment. MoE-HRS further reached F1 and ROC-AUC scores of 0.9823 and 0.9900, respectively, while the second-best CNN model attained 0.9507 and 0.9734. Upon analysis, we determined that the primary factor for this improvement was the inclusion of patient data such as age, family medical history (illnesses of parents and siblings), and lifestyle factors (smoking and alcohol intake), which significantly enhanced the model's performance.

In conclusion, compared to other deep learning models such as GRU, LSTM, Bi-LSTM and SNP2Vec, the MoE-HRS model consistently outperforms baseline models in cancer risk prediction across various cancer types, especially in cases where complex genetic interactions were more difficult to detect using sequential models. The single CNN and single transformer models also performed well, but their results were inferior to the MoE-HRS model, highlighting the benefit of combining both architectures through mixtral of experts. It is important to note that our model is intended as a tool for risk prediction, screening, and decision support rather than serving as a definitive diagnostic instrument; therefore, the tradeoff in precision is balanced by its role in guiding

Table 8. Risk Prediction Results for Prostate Cancer

| | Precision (± Std) | Recall (± Std) | F1 Score (± Std) | ROC-AUC (± Std) |
|---|---|---|---|---|
| PRS-like | 0.5101 (0.005) | 0.4453 (0.007) | 0.4755 (0.006) | 0.5088 (0.006) |
| SVM | **0.6247** (0.004) | 0.5854[a] (0.002) | 0.6044[a] (0.004) | 0.6590[a] (0.003) |
| MLP | 0.5879 (0.003) | 0.4820 (0.004) | 0.5297 (0.005) | 0.6013 (0.004) |
| CNN | 0.5993 (0.004) | 0.5491 (0.006) | 0.5731 (0.005) | 0.6241 (0.005) |
| LSTM | 0.5732 (0.006) | 0.5458 (0.007) | 0.5592 (0.007) | 0.5818 (0.003) |
| Bi-LSTM | 0.5805 (0.007) | 0.5430 (0.009) | 0.5611 (0.008) | 0.5851 (0.005) |
| GRU | 0.5666 (0.003) | 0.5441 (0.004) | 0.5551 (0.003) | 0.5754 (0.006) |
| Transformer | 0.6029 (0.005) | 0.5287 (0.007) | 0.5634 (0.008) | 0.6211 (0.006) |
| SNP2Vec | 0.6146 (0.005) | 0.5056 (0.006) | 0.5473 (0.006) | 0.6309 (0.005) |
| **MoE-HRS** | 0.6153[a] (0.004) | **0.6500** (0.005) | **0.6322** (0.004) | **0.6653** (0.005) |

Results are reported as Mean ± Std over five runs.
Values in bold represent the best result, and those with an asterisk ([a]) represent the second-best result.

further evaluation and intervention. In the meantime, the variability in performance across different cancers underscores the need for ongoing research and potential model refinements tailored to specific types of cancer. Despite the inherent limitations of genomic-based prediction, the value of early risk guidance remains significant; however, translating these predictions into actionable clinical steps represents the next phase of system development.

### 5.6 Feature Interpretation and Causality-Aware Relationships

To identify the genetic, lifestyle and family historical variants most strongly associated with cancer risk prediction across the six cancer types in our study, we employed SHAP analysis [36]. SHAP is a model-agnostic approach that assigns each feature an importance value for a particular prediction based on game theory principles. This method allows us to quantify the contribution of each SNP to the prediction outcome and identify the most influential genetic markers for each cancer type.

For each of the six cancer models, we extracted the top 20 SNPs ranked by their mean absolute SHAP values across the test set, representing the genetic variants with the highest impact on model predictions. Figure 4 illustrates these findings as summary plots where SNPs are ranked by their importance in the prediction model.

For example, as shown in Figure 4(a), among the top SNPs identified for breast cancer, alcohol, and rs1219642_T (FGFR2) emerged as the most influential predictor in our model. This finding aligns with previous GWAS that have consistently identified FGFR2 variants as risk factors for breast cancer [58]. Additionally, rs1219648_A, rs1219651_G, rs34032268_A, rs2912774_T, and rs2981575_G all near FGFR2, were among our top predictors, which have been previously validated in multiple populations [67]. In addition, 10:123340431_GC_G_GC has been demonstrated that associated with breast cancer [47]. The presence of FGFR2 in our top SNPs that located with is particularly noteworthy as these variants have been implicated in estrogen receptor-negative breast cancer subtypes [67], demonstrating our model's ability to capture genetic risk factors across different breast cancer subtypes.

For lung cancer, as illustrated in Figure 4(c), our SHAP analysis highlighted smoke, alcohol and family history are important features that matched with the current related research findings [3]. In addition, rs9272801_T, and rs9272798_C in HLA-DQA1 as the most important SNP, which aligns with extensive evidence linking this locus to lung cancer risk, particularly in smokers [25]. Other prominent SNPs in our model included rs1052576_T and rs933705_T near CASP9 which have been identified in multiple lung cancer GWAS [42].

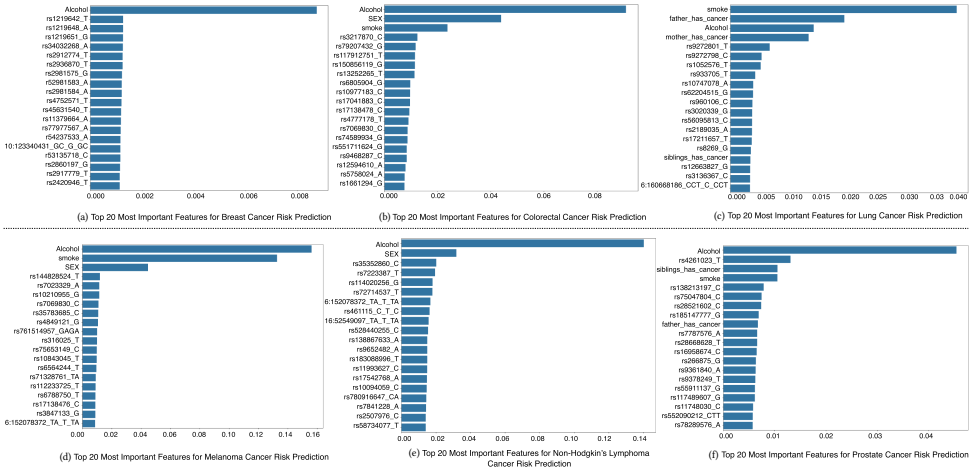


Fig. 4. SHAP-based identification of the top 20 most influential features in cancer risk prediction across six cancer types. The *x*-axis represents the mean absolute SHAP value, reflecting each feature's contribution to the model's predictive performance.

The SNPs identified through our SHAP analysis represent potential biomarkers for cancer risk stratification. The substantial overlap between our findings and previously established genetic associations validates our machine learning approach while also highlighting novel SNPs that merit further investigation. These findings suggest that our MoE-HRS approach, coupled with SHAP-based interpretation, can effectively capture the complex genetic architecture underlying cancer predisposition. The identified SNPs may serve as targets for functional studies to elucidate the biological mechanisms through which they influence cancer risk.

### 5.7 LLM-Powered Healthcare Recommendation

To assess the quality of the healthcare recommendations generated by the LLMs based on our risk predictions, the recommendations are generated using credible sources of medical information. The system provided personalized recommendations for lifestyle changes, preventive measures, and screening schedules. With input from healthcare professionals, the recommendations generated by the LLMs were tailored for clinical medical staff, specifically emphasizing the adjustment of screening intervals based on individual cancer risk.

As depicted in Figure 1(c) and (d), we conduct cancer risk prediction based on the clinical data through the pre-trained MoE-HRS. We categorized risk values as follows: 0.75–1 indicating high cancer risk, 0.6–0.75 suggesting moderate risk, 0.5–0.6 representing mild risk, and below 0.5 denoting low risk. Concurrently, we utilized SHAP to analyze the model, identifying genes associated with the Top@K SNPs linked to risk values. Subsequently, we formulated prompts incorporating both the risk value and associated genes. These prompts were then input into a LLM to generate relevant healthcare recommendations. At the same time, after manual review by clinical medical personnel, we added various cancer-related standard cases to the prompt as reference for the LLMs. As shown in Table 9, prompts were designed based on the high-risk breast cancer patients in the risk prediction model, the relevant Top@3 genes obtained by SHAP analysis, and the text data related to StatPearls (due to the long content, we provide a link to the data resources as a representative example). The corresponding healthcare recommendations were obtained through GPT-4. Specifically, corresponding healthcare recommendations were made through Screening Measures, Self-Examination Techniques, and so on.

Table 9. Case Study for LLM-Driven Healthcare Recommendation

| Prompt | LLM feedback |
|---|---|
| Please provide detailed recommendations for patients at high risk of breast cancer. Focus on the metabolic pathways involved in breast cancer risk, appropriate screening measures, and self-examination techniques.<br>Make sure the information is concise, evidence-based, and only use the given text. Be personalized on the genomic analysis given the top 10 features identified in this patient for risk prediction (Alcohol, rs117912751_T, rs187872526_T, mother_has_cancer, rs191111426_A, rs72806951_G, rs72806949_T, rs34874723_C, rs74487233_C, and rs570718910_G).<br>Provide text extracted from StatPearls, which is available from: https://www.ncbi.nlm.nih.gov/books/NBK430685/ | 1. Metabolic Pathways in Breast Cancer Risk<br>Breast cancer risk is influenced by both genetic and metabolic factors. The following pathways are critical in understanding the patient's genomic risk:<br>1.1. Alcohol Metabolism (SNP: Alcohol, SHAP 2.44e-06): Alcohol consumption is linked to increased estrogen levels, which can promote hormone receptor-positive breast cancers. Recommendation: Limit alcohol intake to no more than one drink per week to reduce estrogen-driven cancer risk.<br>1.2. DNA Repair and Tumor Suppression (SNPs: rs117912751_T, rs191111426_A, rs72806951_G, rs72806949_T): Variants in these SNPs may affect genes involved in DNA repair mechanisms, similar to BRCA mutations.<br>1.3. Inflammatory and Oxidative Stress Pathways (SNPs: rs187872526_T, rs34874723_C, rs74487233_C, rs570718910_G): Chronic inflammation contributes to DNA damage and cancer initiation. Recommendation: Adopt an anti-inflammatory diet rich in omega-3 fatty acids, leafy greens, and antioxidants.<br>1.4. Family History and Genetic Predisposition (SNP: mother_has_cancer, SHAP 6.11e-07): A positive family history doubles to triples breast cancer risk. Recommendation: Consider genetic counseling and possibly BRCA or other relevant genetic testing to refine risk stratification.<br>2. Screening Strategies Based on Genomic Risk<br>Given the high-risk profile, an intensified screening approach is recommended:<br>2.1. Annual Breast MRI (starting age 30–35): MRI is more sensitive for high-risk individuals, especially those with dense breast tissue.<br>2.2. Mammography (yearly from age 35): Digital mammography or tomosynthesis (3D mammography) should be preferred.<br>2.3. Ultrasound (Adjunctive Screening): For dense breast tissue or to evaluate any indeterminate lesions.<br>2.4. Biomarker Screening: Consider serum estrogen, inflammatory markers, and circulating tumor DNA (ctDNA) for monitoring.<br>3. Self-Examination Techniques<br>Although self-examination is no longer the primary screening tool, it remains crucial for self-awareness.<br>3.1. Timing: Perform a self-exam monthly, ideally 1 week after menstruation when breast tissue is less dense.<br>3.2. Visual Inspection: Look for changes in size, shape, nipple inversion, or skin texture (e.g., dimpling).<br>3.3. Palpation: Use the pads of three fingers to apply light, medium, and firm pressure in circular motions. Check the entire breast and axillary area. Examine while lying down, standing, and in front of a mirror.<br>4. Lifestyle and Preventive Measures<br>4.1. Diet: Increase fiber, cruciferous vegetables, and phytoestrogens while reducing processed foods.<br>4.2. Weight Management: Maintain a BMI $< 25$, as obesity is linked to estrogen-driven breast cancer. |

The recommender system is designed to bridge the gap between risk scores from the MoE-HRS algorithm and actionable recommendations for both clinicians and patients. For clinicians, particularly those in primary care, it offers a quick and tailored interpretation of genomic risk data to guide screening and follow-up decisions. For patients, the system demystifies complex medical information, empowering them to better understand potential risk factors and recommended monitoring measures. By serving these two user groups, the LLM-powered recommendations offer a collaborative environment that supports informed decision-making and proactive health management.

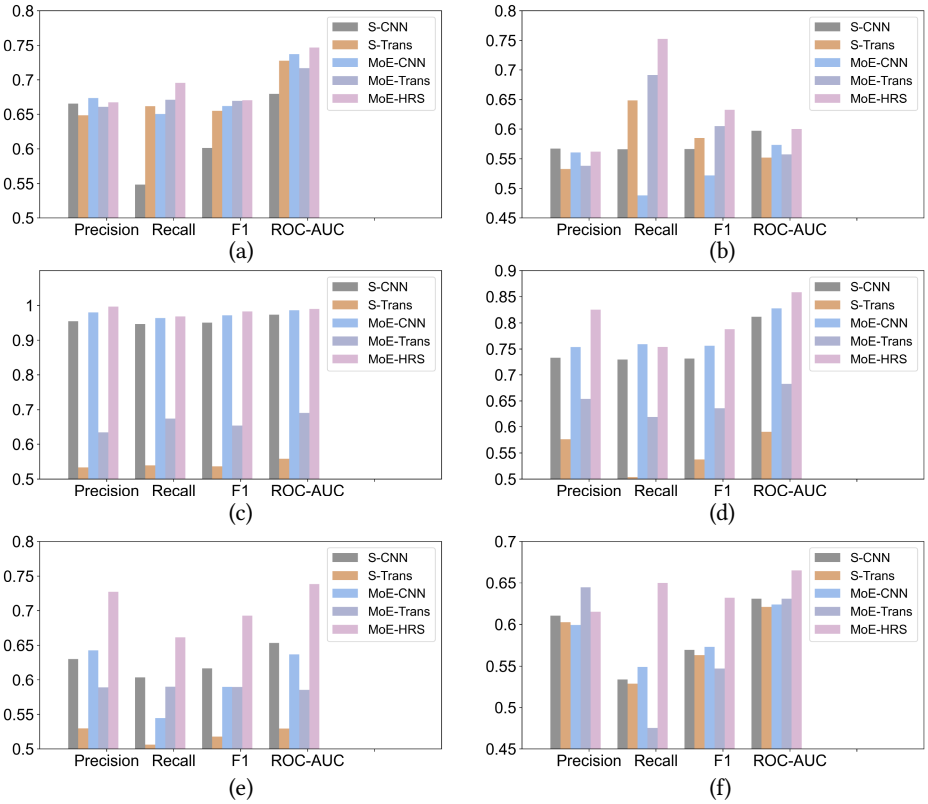


Fig. 5. Ablation experimental results. (a) Ablation study on breast cancer. (b) Ablation study on colorectal cancer. (c) Ablation study on lung cancer. (d) Ablation study on melanoma cancer. (e) Ablation study on NHL cancer. (f) Ablation study on prostate cancer.

### 5.8 Ablation Experiments

To investigate the contribution of each component, we conducted an ablation study by removing the MoE transformer or CNN layers. Through ablation study, we aim to verify the effectiveness of the MoE module in cancer risk prediction. At the same time, we also target at verifying whether our MoE-HRS model is more effective than the Transformer-based MoE model or the CNN-based MoE method. Specifically, we conducted ablation study experiments on single-layer CNN (named S-CNN), CNN-based MoE (named MoE-CNN), single-layer Transformer (named S-Trans), Transformer-based MoE (MoE-Trans), and MoE-HRS. The detailed experimental results are shown in Figure 5.

In general, the model with MoE added performs better than the model without MoE, demonstrating that both components were essential for optimal performance. For example, as shown in Figure 5(d), in melanoma cancer risk prediction, the MoE-CNN has a 4% improvement in Recall and a 3.39% improvement in F1 score compared to the S-CNN model without MoE. At the same time, as shown in Figure 5(c), in lung cancer risk prediction, the MoE-Trans added has an 18.84% improvement in Precision, a 24.96% improvement in Recall, and a 21.81% improvement in F1 score compared to the S-Trans. Simultaneously, as observed in Figure 5, MoE-HRS consistently performed exceptionally well and demonstrated stability in predicting the risk for all six cancers.

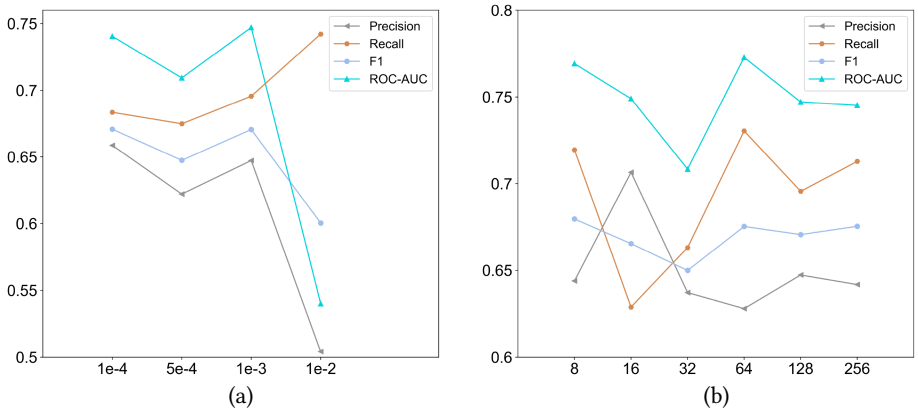


Fig. 6. Hyper-parameters analysis. (a) Learning rate parameters analysis. (b) Embedding dim parameters analysis.

### 5.9 Hyper-Parameters Analysis

In this section, we detail the analysis of hyper-parameters for our proposed model, focusing on their impact on performance metrics such as precision, recall, F1 score and ROC-AUC. We perform analytical experiments focusing on two key parameters: the input embedding dimension and the learning rate. The detailed results are shown in Figure 6.

For learning rate hyper-parameters analysis, we conducted experiments varying the learning rate following 1e-4, 5e-4, 1e-3, and 1e-2. For each dimension, we trained the model and evaluated its performance using relevant metrics such as precision, recall, F1-score, and ROC-AUC with fixed embedding dimension with 128. As shown in Figure 6(a), When the learning rate of the model is 1e-3, Recall, F1 score and ROC-AUC perform best. When the learning rate is 1e-4, precision performs best. When the learning rate is 1e-2, the model effect fluctuates greatly.

For embedding dimension hyper-parameters analysis, we conducted experiments varying the embedding dimension following 8, 16, 32, 64, 128, and 256. For each dimension, we trained the model on our dataset and evaluated its performance using relevant metrics such as precision, recall, F1-score, and ROC-AUC with fixed learning rate of 1e-3. In general, Recall, F1 score, and ROC-AUC performed best when the model embedding dimension was 8 and 64, while precision performed best when the embedding dimension was 128.

## 6 Conclusions and Future Study

### 6.1 Conclusions

In this study, we introduced a novel approach to genetic risk prediction by designing a MoE method based on a Transformer architecture combined with CNN, named MoE-HRS. Our research represents the first attempt to apply the MoE paradigm to genetic risk prediction. By combining a Transformer-based model with a CNN within an MoE framework, this framework leverages both long-range dependencies and local genetic motifs. This dual approach allows for a more comprehensive analysis of genetic sequences, family history records, and lifestyle factors, potentially capturing complex interactions that traditional methods might miss. The gating network in our MoE architecture enables dynamic routing of genetic data to specialized expert networks. As such, our model is able to capture both the contextual and spatial relationships within genetic data, leading to more accurate predictions than conventional methods.

To bridge the gap between risk prediction and practical healthcare applications, we devised a healthcare recommender system powered by an LLM. Prompted by individual risk profiles, the LLMs provides tailored guidance on preventive actions, screening protocols, and lifestyle suggestions to reduce the risk of cancer. This addresses the need for actionable insights following a risk assessment. Notably, this approach demonstrates high predictive accuracy and offers a comprehensive, personalized solution for cancer management along with new possibilities for early intervention and cancer prevention. As such, MoE-HRS not only makes risk predictions more accurate, it also offers practical suggestions to guide clinical decisions and patient self-care. As a holistic solution, MoE-HRS should overcome some of the existing challenges in cancer risk prediction while helping to advance personalized healthcare through early detection and preventive strategies. Our results with six types of cancer indicate a significant improvement in both the accuracy of cancer risk prediction and the relevance of healthcare recommendations. This approach holds promise for enhancing early detection rates and promoting preventive healthcare strategies.

In conclusion, our novel MoE-based approach represents a significant advancement in genetic risk prediction. Our experimental results demonstrate that the proposed MoE method outperforms existing state-of-the-art models in genetic risk prediction tasks. By combining the strengths of transformers and CNN within an MoE framework, we have developed a powerful, interpretable, and adaptable tool for genomic analysis. The improved performance and interpretability of our model have the potential to enhance personalized medicine approaches and deepen our understanding of genetic risk factors. Additionally, the system's ability to provide personalized recommendations could lead to improved patient engagement and potentially better health outcomes in cancer prevention and management.

### 6.2 Future Study

While our novel MoE-based method represents a significant step forward in genetic risk prediction for LLM-driven healthcare recommendation, there remain following three key areas for our further research:

- —XAI Enhancements: Developing more advanced interpretability techniques specifically designed for genetic data could further enhance the model's utility in clinical settings. This could include methods for mapping model predictions to specific genetic pathways or biological processes.
- —Transfer Learning and Pre-training: Investigating the use of transfer learning techniques, where the model is pre-trained on large-scale genetic datasets before fine-tuning on specific prediction tasks.
- —Clinical Validation and Implementation: Focus on rigorous clinical validation of the model across diverse populations and genetic conditions. Additionally, research into the practical implementation of the model in clinical settings, including integration with existing healthcare systems and development of user-friendly interfaces for clinicians, will be essential for translating this technology into real-world impact.

## References

[1] Zeeshan Ahmed, Khalid Mohamed, Saman Zeeshan, and XinQi Dong. 2020. Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database* 2020 (2020), baaa010.

[2] Joanna S. Amberger, Carol A. Bocchini, François Schiettecatte, Alan F. Scott, and Ada Hamosh. 2015. OMIM. org: Online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* 43, Database issue (2015), D789–D798.

[3] Elisa V. Bandera, Jo. L. Freudenheim, and John E. Vena. 2001. Alcohol consumption and lung cancer: A review of the epidemiologic evidence. *Cancer Epidemiology Biomarkers & Prevention* 10, 8 (2001), 813–821.
[4] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. 2022. On Pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*. Association for Computing Machinery, New York, NY, USA, 507–556. Retrieved from https://dl.acm.org/doi/abs/10.1145/3501714
[5] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. 2024. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a Cancer Journal for Clinicians* 74, 3 (2024), 229–263.
[6] Samuel Cahyawijaya, Tiezheng Yu, Zihan Liu, Tiffany T. W. Mak, Xiaopu Zhou, Nancy Y. Ip, and Pascale Fung. 2022. SNP2Vec: Scalable self-supervised pre-training for genome-wide association study. arXiv:2204.06699. Retrieved from https://arxiv.org/abs/2204.06699
[7] Nilanjan Chatterjee, Jianxin Shi, and Montserrat García-Closas. 2016. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nature Reviews Genetics* 17, 7 (2016), 392–406.
[8] Kyunghyun Cho. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv:1406.1078. Retrieved from https://arxiv.org/abs/1406.1078
[9] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 3 (1995), 273–297.
[10] Michael Elgart, Genevieve Lyons, Santiago Romero-Brufau, Nuzulul Kurniansyah, Jennifer A. Brody, Xiuqing Guo, Henry J. Lin, Laura Raffield, Yan Gao, Han Chen, et al. NHLBI's trans-omics in precision medicine (TOPMed) consortium. 2022. Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Communications Biology* 5, 1 (2022), 856.
[11] Mitchell H. Gail, Louise A. Brinton, David P. Byar, Donald K. Corle, Sylvan B. Green, Catherine Schairer, and John J. Mulvihill. 1989. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* 81, 24 (1989), 1879–1886.
[12] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional LSTM networks for improved phoneme classification and recognition. In *International Conference on Artificial Neural Networks*. Springer, 799–804.
[13] Jiafeng Guo, Yinqiong Cai, Keping Bi, Yixing Fan, Wei Chen, Ruqing Zhang, and Xueqi Cheng. 2024. CAME: Competitively learning a mixture-of-experts model for first-stage retrieval. *ACM Transactions on Information Systems* 43, 2 (2024), 1–25.
[14] Kairui Guo, Mengjia Wu, Zelia Soo, Yue Yang, Yi Zhang, Qian Zhang, Hua Lin, Mark Grosser, Deon Venter, Guangquan Zhang, et al. 2023. Artificial intelligence-driven biomedical genomics. *Knowledge-Based Systems* 279 (2023), 110937.
[15] Jennifer S. Haas, Celia P. Kaplan, Steven E. Gregorich, Eliseo J. Pérez-Stable, and Genevieve Des Jarlais. 2004. Do physicians tailor their recommendations for breast cancer risk reduction based on patient's risk? *Journal of General Internal Medicine* 19, 4 (2004), 302–309.
[16] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[17] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. 2018. Artificial intelligence in radiology. *Nature Reviews Cancer* 18, 8 (2018), 500–510.
[18] Shigao Huang, Jie Yang, Simon Fong, and Qi Zhao. 2020. Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters* 471 (2020), 61–71.
[19] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. 2020. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2907–2916.
[20] Guochong Jia, Yingchang Lu, Wanqing Wen, Jirong Long, Ying Liu, Ran Tao, Bingshan Li, Joshua C. Denny, Xiao-Ou Shu, and Wei Zheng. 2020. Evaluating the utility of polygenic risk scores in identifying high-risk individuals for eight common cancers. *JNCI Cancer Spectrum* 4, 3 (2020), pkaa021.
[21] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. arXiv:2401.04088. Retrieved from https://arxiv.org/abs/2401.04088
[22] Linda Kachuri, Rebecca E. Graff, Karl Smith-Byrne, Travis J. Meyers, Sara R. Rashkin, Elad Ziv, John S. Witte, and Mattias Johansson. 2020. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nature Communications* 11, 1 (2020), 6084.
[23] Robert J. Klein, Emily Vertosick, Dan Sjoberg, David Ulmert, Ann-Charlotte Rönn, Christel Häggström, Elin Thysell, Göran Hallmans, Anders Dahlin, Pär Stattin, et al. 2022. Prostate cancer polygenic risk score and prediction of lethal prostate cancer. *NPJ Precision Oncology* 6, 1 (2022), 25.
[24] Robert J. Klein, Caroline Zeiss, Emily Y. Chew, Jen-Yue Tsai, Richard S. Sackler, Chad Haynes, Alice K. Henning, John Paul SanGiovanni, Shrikant M. Mane, Susan T. Mayne, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)* 308, 5720 (2005), 385–389.
[25] Takashi Kohno, Hideo Kunitoh, Sachiyo Mimaki, Kouya Shiraishi, Aya Kuchiba, Seiichiro Yamamoto, and Jun Yokota. 2011. Contribution of the TP53, OGG1, CHRNA3, and HLA-DQA1 genes to the risk for lung squamous cell carcinoma. *Journal of Thoracic Oncology* 6, 4 (2011), 813–817.

[26] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* 13 (2015), 8–17.

[27] Samuel A. Lambert, Gad Abraham, and Michael Inouye. 2019. Towards clinical utility of polygenic risk scores. *Human Molecular Genetics* 28, R2 (2019), R133–R142.

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 11 (1998), 2278–2324.

[29] Cathryn M. Lewis and Evangelos Vassos. 2020. Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine* 12, 1 (2020), 44.

[30] Ruowang Li, Yong Chen, Marylyn D. Ritchie, and Jason H. Moore. 2020. Electronic health records and polygenic risk scores for predicting disease risk. *Nature Reviews Genetics* 21, 8 (2020), 493–502.

[31] Javier Louro, Margarita Posso, Michele Hilton Boon, Marta Román, Laia Domingo, Xavier Castells, and María Sala. 2019. A systematic review and quality assessment of individualised breast cancer risk prediction models. *British Journal of Cancer* 121, 1 (2019), 76–85.

[32] Jie Lu, Junyu Xuan, Guangquan Zhang, and Xiangfeng Luo. 2018. Structural property-aware multilayer network embedding for latent factor analysis. *Pattern Recognition* 76 (2018), 228–241.

[33] Kezhi Lu, Kuo Yang, Hailong Sun, Qian Zhang, Qiguang Zheng, Kuan Xu, Jianxin Chen, and Xuezhong Zhou. 2023. SympGAN: A systematic knowledge integration system for symptom–gene associations network. *Knowledge-Based Systems* 276 (2023), 110752.

[34] Kezhi Lu, Qian Zhang, Danny Hughes, Guangquan Zhang, and Jie Lu. 2024. AMT-CDR: A deep adversarial multi-channel transfer network for cross-domain recommendation. *ACM Transactions on Intelligent Systems and Technology* 15, 4 (2024), 1–26.

[35] Kezhi Lu, Qian Zhang, Guangquan Zhang, and Jie Lu. 2023. BERT-RS: A neural personalized recommender system with BERT. In *15th International FLINS Conference on Machine Learning, Multi Agent and Cyber Physical Systems (FLINS '22)*. World Scientific, 390–397.

[36] Scott Lundberg. 2017. A unified approach to interpreting model predictions. arXiv:1705.07874. Retrieved from https://arxiv.org/abs/1705.07874

[37] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2, 10 (2018), 749–760.

[38] Alicia R. Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M. Neale, and Mark J. Daly. 2019. Current clinical use of polygenic scores will risk exacerbating health disparities. *Nature Genetics* 51, 4 (2019), 584–591.

[39] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P. Tyrer, Ting-Huei Chen, Qin Wang, Manjeet K. Bolla, et al. 2019. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *The American Journal of Human Genetics* 104, 1 (2019), 21–34.

[40] Luke McGeoch, Catherine L. Saunders, Simon J. Griffin, Jon D. Emery, Fiona M. Walter, Deborah J. Thompson, Antonis C. Antoniou, and Juliet A. Usher-Smith. 2019. Risk prediction models for colorectal cancer incorporating common genetic variants: A systematic review. *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 28, 10 (2019), 1580–1593.

[41] Catherine Meads, Ikhlaaq Ahmed, and Richard D. Riley. 2012. A systematic review of breast cancer incidence risk prediction models with meta-analysis of their performance. *Breast Cancer Research and Treatment* 132, 2 (2012), 365–377.

[42] Jae Yong Park, Jung Min Park, Jin Sung Jang, Jin Eun Choi, Kyung Mee Kim, Sung Ick Cha, Chang Ho Kim, Young Mo Kang, Won Kee Lee, Sin Kam, et al. 2006. Caspase 9 promoter polymorphisms and risk of primary lung cancer. *Human Molecular Genetics* 15, 12 (2006), 1963–1971.

[43] Giulia Pasello, Alberto Pavan, Ilaria Attili, Alberto Bortolami, Laura Bonanno, Jessica Menis, PierFranco Conte, and Valentina Guarneri. 2020. Real world data in the era of immune checkpoint inhibitors (ICIs): increasing evidence and future applications in lung cancer. *Cancer Treatment Reviews* 87 (2020), 102031.

[44] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I. Furlong. 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* 48, D1 (2020), D845–D855.

[45] International Schizophrenia Consortium; Shaun M. Purcell, Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O'Donovan, Patrick F. Sullivan, and Pamela Sklar. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 7256 (2009), 748–752.

[46] Noa Rappaport, Michal Twik, Inbar Plaschkes, Ron Nudel, Tsippi Iny Stein, Jacob Levitt, Moran Gershoni, C. Paul Morrey, Marilyn Safran, and Doron Lancet. 2017. MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research* 45, D1 (2017), D877–D887.

[47] Sara R. Rashkin, Rebecca E. Graff, Linda Kachuri, Khanh K. Thai, Stacey E. Alexeeff, Maruta A. Blatchins, Taylor B. Cavazos, Douglas A. Corley, Nima C. Emami, Joshua D. Hoffman, et al. 2020. Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nature Communications* 11, 1 (2020), 4423.

[48] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

[49] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature* 323, 6088 (1986), 533–536.

[50] David L. Sackett, William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. Evidence based medicine: What it is and what it isn't. B*mj* 312, 7023 (1996), 71–72.

[51] Qusai Shambour and Jie Lu. 2015. An effective recommender system by unifying user and item trust information for B2B applications. *Journal of Computer and System Sciences* 81, 7 (2015), 1110–1126.

[52] Edward H. Shortliffe and Bruce G. Buchanan. 1975. A model of inexact reasoning in medicine. *Mathematical Biosciences* 23, 3–4 (1975), 351–379.

[53] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.

[54] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. 2015. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of Middle and old age. *PLoS Medicine* 12, 3 (2015), e1001779.

[55] Martin C. Tammemägi, Hormuzd A. Katki, William G. Hocking, Timothy R. Church, Neil Caporaso, Paul A. Kvale, Anil K. Chaturvedi, Gerard A. Silvestri, Tom L. Riley, John Commins, et al. 2013. Selection criteria for lung-cancer screening. *The New England Journal of Medicine* 368, 8 (2013), 728–736.

[56] Ali Torkamani, Nathan E. Wineinger, and Eric J. Topol. 2018. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* 19, 9 (2018), 581–590.

[57] Khoa A. Tran, Olga Kondrashova, Andrew Bradley, Elizabeth D. Williams, John V. Pearson, and Nicola Waddell. 2021. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine* 13, 1 (2021), 152–117.

[58] Miriam S. Udler, Kerstin B. Meyer, Karen A. Pooley, Eric Karlins, Jeffery P. Struewing, Jinghui Zhang, David R. Doody, Stewart MacArthur, Jonathan Tyrer, Paul D. Pharoah, et al. 2009. FGFR2 variants and breast cancer risk: Fine-scale mapping using African American studies and analysis of chromatin conformation. *Human Molecular Genetics* 18, 9 (2009), 1692–1703.

[59] A. Vaswani. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

[60] Peter M. Visscher, Loic Yengo, Nancy J. Cox, and Naomi R. Wray. 2021. Discovery and implications of polygenicity of common diseases. *Science (New York, N.Y.)* 373, 6562 (2021), 1468–1473.

[61] Hannah Wand, Samuel A. Lambert, Cecelia Tamburro, Michael A. Iacocca, Jack W. O'Sullivan, Catherine Sillari, Iftikhar J. Kullo, Robb Rowley, Jacqueline S. Dron, Deanna Brockman, et al. 2021. Improving reporting standards for polygenic scores in risk prediction studies. *Nature* 591, 7849 (2021), 211–219.

[62] Ying Wang, Kristin Tsuo, Masahiro Kanai, Benjamin M. Neale, and Alicia R. Martin. 2022. Challenges and opportunities for developing more generalizable polygenic risk scores. *Annual Review of Biomedical Data Science* 5, 1 (2022), 293–320.

[63] Naomi R. Wray, Tian Lin, Jehannine Austin, John J. McGrath, Ian B. Hickie, Graham K. Murray, and Peter M. Visscher. 2021. From basic science to clinical application of polygenic risk scores: A primer. *JAMA Psychiatry* 78, 1 (2021), 101–109.

[64] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *8th cCF International Conference on Natural Language Processing and Chinese Computing (NLPCC '19)*. Springer, 563–574.

[65] Zijun Yao, Bin Liu, Fei Wang, Daby Sow, and Ying Li. 2023. Ontology-aware prescription recommendation in treatment pathways using multi-evidence healthcare data. *ACM Transactions on Information Systems* 41, 4 (2023), 1–29.

[66] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology* 15, 2 (2024), 1–38.

[67] Yihan Zhao, Di Wu, Danli Jiang, Xiaoyu Zhang, Ting Wu, Jing Cui, Min Qian, Jean Zhao, Steffi Oesterreich, Wei Sun, et al. 2020. A sequential methodology for the rapid identification and characterization of breast cancer-associated functional SNPs. *Nature Communications* 11, 1 (2020), 3340.

[68] Siqiong Zhou, Upala J. Islam, Nicholaus Pfeiffer, Imon Banerjee, Bhavika K. Patel, and Ashif S. Iquebal. 2023. SCGAN: Sparse CounterGAN for counterfactual explanations in breast cancer prediction. *IEEE Transactions on Automation Science and Engineering* 21, 3 (2023), 2264–2275.

Received 29 September 2024; revised 16 May 2025; accepted 25 May 2025