A Fair and Personalized Dementia Prediction Framework Using Longitudinal and Demographic Data from South Korea

Hong-Woo Chun[1], Lee-Nam Kwon[1], Hyeonho Shin[1], SungWha Hong[1] and Jae-Min Lee[1,*]

[1]*Future Information Analysis Center, Korea Institute of Science and Technology Information, 66 Hoegi-ro, Dongdaemun-gu, Seoul, 02456, South Korea*

Abstract

Early prediction of dementia is a critical public health challenge, yet conventional machine learning models often treat all patients as a single, uniform population. This approach overlooks subtle clinical differences between individuals and can lead to biased predictions that disproportionately affect specific demographic groups. This study proposes a novel framework that leverages the power of a Large Language Model (LLM) to build a fair and personalized dementia prediction system. While traditional methods required separate modeling for men and women, the LLM, thanks to its reasoning capabilities, can perform customized predictions using all data without the need for such separate modeling. We show that providing an LLM with specific demographic context, such as sex, leads to more nuanced and accurate predictive results than a generic, non-contextual prompt. This approach demonstrates that LLMs can be a powerful tool for developing personalized medical AI systems that respect individual differences and reduce algorithmic bias.

Keywords

LLM-based Dementia Early Prediction, Personalization, Fairness and Bias in Medical AI

1. Introduction

Dementia is a growing public health challenge worldwide, with global estimates predicting a sharp increase in patient numbers – a trend that is particularly pronounced in South Korea where long-term health databases offer unique insights into population trends [1]. Traditional dementia prediction models—typically based on statistical learning from imaging, genetic markers, or cross-sectional clinical data—tend to pool all patients into a single homogeneous group, often overlooking crucial demographic differences that influence both disease onset and progression [2] [3]. Such "one-size-fits-all" paradigm has led to models that perform well in aggregate but tend to be biased against underrepresented subpopulations, for example, failing to capture sex-specific differences in Alzheimer's disease risk and presenting lower predictive accuracy for women [4]. Recent research has increasingly emphasized that fairness and bias in medical artificial intelligence must be addressed throughout the model development lifecycle – from data preprocessing to algorithmic design – to avoid perpetuating existing health disparities [5]. The advent of Large Language Models (LLMs) offers a transformative opportunity in this context. Unlike conventional machine learning approaches, LLMs possess advanced reasoning and generative capabilities that allow them to integrate unstructured data and contextual variables, such as explicit demographic details, to deliver more personalized and equitable predictions [1]. This work presents a novel dementia prediction framework that leverages an LLM with a demographically-aware prompting strategy, enabling it to generate context-sensitive predictions by integrating longitudinal clinical data with critical demographics. By doing so, our framework not only improves overall prediction accuracy but also directly mitigates bias, ensuring that individualized risk factors are considered rather

Identity-Aware AI 2025, October 25, 2025, Bologna, ITALY

*Corresponding author.

hw.chun@kisti.re.kr (H. Chun); ynkwon@kisti.re.kr (L. Kwon); shinhh9554@kisti.re.kr (H. Shin); shong@kisti.re.kr (S. Hong); jmlee@kisti.re.kr (J. Lee)

0000-0002-9584-7065 (H. Chun); 0000-0002-3503-7669 (L. Kwon); 0000-0001-7130-152X (H. Shin); 0009-0001-2552-9379 (S. Hong); 0000-0002-4011-987X (J. Lee)

than smoothed over in aggregate models [4] [3].

2. Related work

In this section, we briefly review research trends in dementia prediction and then focus on recent advances in fairness and bias mitigation within medical AI, particularly in the era of LLM-driven clinical decision-making.

2.1. Research Trends in Dementia Prediction

This section reviews two sets of research: one focusing on advancements in dementia prediction and the other on fairness and bias mitigation in medical AI, with special attention to generative and LLM-driven approaches.

2.2. Research Trends in Dementia Prediction

Early approaches to dementia prediction mainly relied on single-point in time laboratory assessments or neuroimaging data such as MRI and PET scans, with genetic markers like the ApoE4 allele providing additional risk information [6]. Although these methods have delivered foundational insights into disease etiology, they typically fall short in capturing the progressive and multifactorial nature of dementia. More recent studies have increasingly utilized longitudinal health records—such as those available from South Korea's National Health Information Database—to track changes in clinical parameters over time, thereby significantly improving risk prediction [6]. However, these models largely rely on traditional approaches such as support vector machines or multilayer perceptrons and struggle to adapt to individual-level variations, often resulting in differential predictive performance across demographic groups [2]. This limitation motivates the emerging interest in advanced deep learning techniques, such as LLM-based frameworks, which offer intrinsic mechanisms for assimilating heterogeneous data and can be tailored via context-specific prompting to address inherent disparities [1].

2.3. Fairness and Bias in Medical AI

Fairness and bias mitigation have become critical as medical AI systems are deployed in clinical decision-making, especially given that biased training data and model architectures can amplify existing disparities in healthcare outcomes [3] [5]. For instance, models trained on predominantly Western or white populations often underperform when applied to minority groups, resulting in lower diagnostic accuracy and delayed treatment for underserved patients [4] [2]. Recent research on medical AI has explored a spectrum of techniques to decrease such bias. Approaches such as adversarial debiasing, counterfactual fairness testing, output recalibration, and threshold adjustment have been applied to reduce disparate impacts by ensuring that model predictions remain accurate and equitable when evaluated separately across demographic subgroups [2]. These technical strategies are often paired with explainability frameworks—which utilize methods such as LIME, SHAP, or attention-based visualizations—to provide transparent insight into the model's decision-making process. Such interpretability is essential for clinical trust and to diagnose hidden sources of bias because it allows stakeholders to evaluate how patient-specific factors such as sex, race, or socioeconomic status influence predictions [3].

The integration of fairness-aware methods with LLMs is an emerging trend in healthcare AI. Recent studies have demonstrated that providing LLMs with explicit demographic context through tailored prompting strategies leads to substantially improved prediction accuracy and fairness compared to generic, noncontextual approaches [4]. In one example, researchers used demographic perturbation and prompt engineering to identify and mitigate bias in diagnostic outputs, highlighting that even state-of-the-art LLMs may exhibit disparities unless calibrated with subgroup-specific information

[7]. Moreover, frameworks employing federated learning and fair representation learning have shown promise in harnessing diverse datasets without compromising patient privacy, reinforcing the necessity of interdisciplinary strategies that span technical, ethical, and regulatory domains [1] [3].

In addition to these technical advances, the literature increasingly calls for standardized fairness metrics and comprehensive evaluation frameworks. Benchmarking tools such as MEDFAIR and DiversityMedQA have been introduced to assess algorithmic fairness under various conditions, providing quantifiable measures of equity across racial, sex, and socioeconomic dimensions [2] [5]. These efforts are complemented by policy-driven initiatives which seek to embed fairness considerations into the broader medical AI regulatory landscape, an evolution that is essential to promote both accountability and trust in clinical settings [2] [3].

By synthesizing these lines of research, our study adopts an LLM-based approach that directly confronts the limitations of traditional one-size-fits-all models. Our framework leverages demographically-aware prompting to provide personalized dementia risk predictions, thereby simultaneously enhancing overall accuracy and reducing bias. This work contributes to the growing body of literature that seeks to advance not only the technical performance but also the ethical underpinnings of AI applications in healthcare [8]. Future research directions include expanding evaluations to additional demographic subgroups and integrating state-of-the-art fairness mitigation techniques that continue to emerge in response to evolving clinical standards [7].

In summary, revised research in dementia prediction now not only emphasizes the importance of longitudinal and heterogeneous data integration but also necessitates a shift toward fairness-focused AI methodologies. By harnessing the next-generation capabilities of LLMs combined with rigorous bias auditing and explainability measures, our approach promises a fairer, more individualized method for early dementia risk assessment—one that is responsive to the complex interplay of demographic factors and clinical indicators [1] [3].

3. Methodology

This study applied a novel LLM-based methodology to analyze bias in dementia prediction and prove the need for a personalized approach. Instead of training distinct machine learning models, we utilized a single LLM with a flexible prompting strategy to generate predictive insights.

3.1. Dataset

This study uses the National Health Information Database (NHID) of the National Health Insurance Service in South Korea, specifically the NHIS-SC DB, to predict the risk of developing dementia [9] [10]. The NHIS-SC DB is a large-scale cohort of one million individuals over the age of 60, randomly sampled from the total population. This database contains comprehensive health-related information for these individuals from 2002 to 2013, including health insurance eligibility, medical treatments, diagnoses, health check-up results, and long-term care insurance. We used the following key sub-databases:

- **PIE-DB (Population-based Individualized Eligibility-DB)**: Contains demographic information (sex, age, residential area) and health insurance eligibility.
- **MT-DB (Medical Treatment-DB)**: Provides detailed records of outpatient visits, hospitalizations, and prescriptions, coded using the Korean Standard Classification of Diseases (KCD), which is based on the International Classification of Diseases (ICD).
- **GHE-DB (General Health Examination-DB)**: Includes results from biennial health check-ups, such as blood pressure, BMI, blood sugar levels, and cholesterol.
- **MCI-DB (Medical care institution DB)**: Comprises data such as type of medical institution, area and installation period, number of hospital beds, number of doctors, and equipment availability status.

- **LCI-DB (Long-term care insurance DB)**: Encompasses long-term care application and decision results, opinions of doctors, such as an examination of recognized necessity, long-term care facility data.

3.2. LLM-Based Prediction System Architecture and Experimental Setup

The LLM-based dementia prediction system proposed in this study has a fundamentally different architecture from traditional machine learning models. Instead of individually training multiple models, it leverages the powerful reasoning capabilities of a single LLM to derive personalized predictions through data preprocessing and prompt engineering.

System Architecture The overall flow of our LLM-based prediction system is illustrated below. The architecture is modular and designed to maximize the LLM's natural language understanding and generative capabilities.

Step 1. Data Preprocessing Stage:

- **Input Data**: The input for this stage is the raw data from the NHIS-SC DB. This data consists of various forms of tabular data, including patient medical records, diagnoses, prescription details, and health check-up results.
- **Conversion Process**: While traditional machine learning models would directly use this tabular data as a numerical feature vector, our system maximizes the LLM's natural language processing capabilities by converting this data into narrative text.
- **Medical History Summarization**: The patient's diagnoses (e.g., hypertension, diabetes, depression) and their onset times are organized into a chronological narrative.
- **Health Check-up Records**: Changes in regular health check-up metrics like blood pressure, BMI, blood sugar, and cholesterol are summarized into sentences that describe trends (e.g., "Blood pressure has been consistently rising over the past three years").
- **Demographic Information**: Demographic details such as sex, age, and residential area are explicitly stated.
- **Output**: After this stage, all medical information for each patient is generated as a single, unified text block that is easily understandable by the LLM.

Step 2. LLM-Based Prediction Stage:

- **Prompt Design**: This stage is crucial for the LLM's performance. We conducted our experiments using two distinct prompting strategies. The core idea is to test how the presence of explicit demographic context influences the LLM's predictive reasoning.
- **Integrated Prompt**: This prompt provided the LLM with a patient's structured medical history without any specific demographic context. The instruction was simply "Based on the following medical history, predict the likelihood of dementia and explain your reasoning."
- **Demographically-Aware Prompt**: This prompt included the same medical history but with explicit demographic information. For example, "Based on the following medical history for a 65-year-old female, predict the likelihood of dementia and explain your reasoning."
- **LLM's Reasoning**: The LLM does not simply classify the input text; instead, it performs complex reasoning based on its vast, pre-trained medical knowledge base. For instance, when it receives a demographically-aware prompt, it engages in an internal reasoning process akin to asking, "What are the common risk factors for dementia in females?" This is a qualitatively different process from the MLP model, which merely learns a weight-based function. The LLM can draw upon its knowledge that "depression has a greater impact on dementia in women" to make its prediction and provide a detailed explanation that goes beyond simple correlation.

Step 3. Output Analysis and Evaluation Stage:

- **LLM's Output**: The LLM generates a prediction (e.g., "High likelihood of dementia") along with a detailed natural language explanation for its reasoning.
- **Evaluation Methods**: We evaluated the LLM's predictions using two key metrics:
- **Predictive Consistency**: This is a quantitative measure of how well the LLM's prediction aligns with the actual diagnosis in the test set.
- **Reasoning Nuance**: A panel of medical experts (N=5) qualitatively evaluated the LLM's explanations on a 1 (generic) to 5 (highly specific and contextual) scale. This qualitative assessment helps us understand how well the LLM internalizes and applies sex-specific risk factors.

This architecture enables us to overcome the limitations of traditional modeling and demonstrate that a single LLM can achieve personalized predictions while mitigating demographic bias.

Experimental Setup

In this study, we utilized the gemini-2.5-flash-preview-05-20 model for our experiments, chosen for its strong performance in medical text analysis and its capacity for complex reasoning. To ensure stable and consistent output suitable for an academic study, we meticulously tuned the LLM's parameters. The temperature was set to a low value of 0.2 to minimize creative or speculative responses and ensure consistent, fact-based predictions. The max_tokens was set to 2000, allowing the LLM to provide detailed, well-reasoned explanations for its predictions. Each patient's preprocessed medical information text block was capped at 10,000 characters to optimize its fit within the LLM's context window.

4. Experiments and Results

Our experiments compared the predictive performance of a traditional Machine Learning (MLP) framework against a Large Language Model (LLM) utilizing a demographically-aware prompting strategy. The results demonstrate the LLM's ability to achieve more equitable and enhanced performance within a single, unified system, overcoming the structural limitations inherent in traditional models.

4.1. Sex Bias and Limitations of Traditional Machine Learning Models

Traditional machine learning models often treat all patients as a single, uniform population, overlooking subtle clinical differences. This "one-size-fits-all" approach, when applied to dementia prediction, typically results in biased predictions. To address the bias, we analyzed the necessity of separate modeling. We compared the feature importance ranking of risk factors identified by models trained separately for men and women, as illustrated in Table 1 and Table 2. This analysis confirmed that dementia risk factors differ significantly by sex:

- The male prediction model prioritized factors such as Parkinson's disease and other mental disorders due to brain damage.
- In contrast, the female prediction model gave higher priority to factors related to cerebral infarction, depressive episode, and cerebrovascular conditions.

This clear divergence confirms that information from one sex can act as 'noise' in the prediction model for the other sex. Therefore, to address this demographic bias, traditional methods required separate modeling for men and women.

Even after implementing sex-separated modeling, the MLP models exhibited persistent performance disparity:

- The F-measure for the male prediction model was 78.3%.
- The F-measure for the female prediction model was 72.8%.

This result indicated that the traditional approach, even when meticulously separated, still struggled to capture the necessary nuances, leading to significantly lower performance for women (72.8%) compared to men (78.3%).

Table 1
Performance Metrics of Sex-Separated MLP Models (Men)

	Men's RF Top 10	Men + Women RF Rank
1	Parkinson's disease	19
2	Other mental disorders due to brain damage and dysfunction and to physical disease	22
3	Paraplegia and tetraplegia	50
4	Vitamin D deficiency	73
5	Schizophrenia	10
6	Eating disorders	52
7	Other disorders of nervous system, NEC	32
8	Chronic kidney disease	98
9	Acute nephritic syndrome	94
10	Status epilepticus	51

Table 2
Performance Metrics of Sex-Separated MLP Models (Women)

	Women's RF Top 10	Women + Men RF Rank
1	Cerebral infarction	4
2	Other degenerative diseases of nervous system, NEC	8
3	Paraplegia and tetraplegia	50
4	Delirium, not induced by alcohol and	12
o	ther psychoactive substances	
5	Inflammatory disease of uterus, except cervix	89
6	Unspecified urinary incontinence	92
7	Other disorders of pancreatic internal secretion	31
8	Other mental disorders due to brain damage and dysfunction and to physical disease	22
9	Vascular syndromes of brain in cerebro-vascular diseases	17
10	Depressive episode	2

4.2. Effectiveness of LLM-Based Personalized Prediction

In sharp contrast to the traditional approach, the LLM-based methodology achieved superior predictive accuracy for both sexes using a single, unified framework, eliminating the need for complex, separate model training. The LLM (specifically the gemini-2.5-flash-preview-05-20 model) leverages its powerful reasoning capabilities through a Demographically-Aware Prompt strategy. By explicitly providing demographic context (e.g., sex and age) alongside the medical history, the LLM integrates sex-specific risk factors into its internal reasoning process. For example, the LLM can draw upon its extensive knowledge that "depression has a greater impact on dementia in women" to make a more nuanced prediction. When evaluated using the F-measure (for comparative purposes):

- The F-measure for the male prediction model was 79.1%.
- The F-measure for the female prediction model was 82.5%

These results demonstrate that the LLM successfully surpassed the performance of the traditional MLP sex-separated models for both men (79.1% > 78.3%) and women (82.5% > 72.8%) (Table 3). Critically, the LLM achieved its highest performance in the female subgroup (82.5%), effectively mitigating the substantial bias and lower accuracy that plagued the traditional approach for women (72.8%). This

Table 3
Evaluation Comparison

Approach	Model Type and Strategy	Sex	Performance (F-measure, %)
Traditional ML (MLP)	Sex-separated Model	Male	78.3
		Female	72.8
LLM	Unified framework	Male	79.1
	(Demographically-Aware Prompt)	Female	82.5

validates the LLM as an innovative methodology for developing personalized medical AI systems that respect individual differences and reduce algorithmic bias without requiring separate modeling.

5. Conclusion

This study proposed a new methodology for fair and personalized dementia prediction models by leveraging the capabilities of a Large Language Model. By using a demographically-aware prompting strategy with patient data from the NHIS-SC DB, we overcame the limitations of the one-size-fits-all approach that traditional models suffer from. Our research shows that in sensitive medical fields like dementia prediction, it is essential to move away from a universal approach and adopt a personalized one that respects the unique characteristics of each individual, using the LLM's reasoning abilities.

6. Future Plan

In this study, the gemini-2.5-flash-preview-05-20 model showed encouraging performance in medical text analysis and complex reasoning, effectively supporting our clinical data–driven predictive modeling tasks. Nonetheless, to fully understand its relative strengths and limitations, it is necessary to compare its performance with other state-of-the-art LLMs in future work.

The proposed comparative analysis will include widely used state-of-the-art LLMs such as GPT-4, LLaMA, and Claude, in addition to Gemini. Each of these models differs in terms of architecture, training data sources, and reasoning capabilities, which are expected to result in varying performance in processing clinical narratives combined with demographic information. This study aims to evaluate how these models perform in terms of personalized dementia prediction—particularly in aspects of fairness and accuracy.

Future experiments will be conducted by applying the same longitudinal health dataset and unified prompt engineering strategies across all models. The evaluation will incorporate a comprehensive set of quantitative metrics, including F1-score, accuracy, recall, and precision. Furthermore, fairness assessments will be performed across demographic subgroups such as ethnicity, sex, and age to measure each model's bias mitigation effectiveness. We will also consider analyzing the impact of intersectional identities (e.g., age-gender combinations) on predictions and fairness in future work.

In addition, explainability—a core focus of our research—will be systematically assessed. Expert panels will conduct qualitative evaluations of the generated model explanations, and quantitative measures such as the Explanation Satisfaction Scale (ESS) will be applied. To further contextualize LLM explanation quality, we plan to compare these outputs against traditional explainable AI (XAI) approaches such as SHAP and LIME, enabling an objective analysis of practical interpretability in clinical decision-making. The results of this comparative work are expected to provide essential evidence for developing an optimal LLM-based dementia prediction and explanation framework for real-world clinical settings. By systematically considering model performance, fairness, and ethical factors, we aim to enhance the fair and personalized use of AI in healthcare and offer practical guidelines for LLM selection. Moreover, the findings will serve as a foundation for extending this research framework to diverse applications in medical AI and beyond.

Acknowledgments

This research was supported by Korea Institute of Science and Technology Information(KISTI).(No. K25L4M2C4)

Declaration on Generative AI

During the preparation of this work, the authors used Gemini, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. A. Aljohani, J. Hou, S. Kommu, X. Wang, A comprehensive survey on the trustworthiness of large language models in healthcare, ArXiv abs/2502.15871 (2025).

[2] S. V. Chinta, Z. Wang, X. Zhang, T. D. Viet, A. Kashif, M. A. Smith, W. Zhang, Ai-driven healthcare: A survey on ensuring fairness and mitigating bias, ArXiv abs/2407.19655 (2024).

[3] R. J. Chen, J. J. Wang, D. F. K. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, F. Mahmood, Algorithmic fairness in artificial intelligence for medicine and healthcare, Nature Biomedical Engineering 7 (2023) 719–742.

[4] G. Franklin, R. Stephens, M. Piracha, S. Tiosano, F. Lehouillier, R. Koppel, P. Elkin, The sociodemographic biases in machine learning algorithms: A biomedical informatics perspective, Life 14 (2024) 652.

[5] M. Omar, V. Sorin, R. Agbareia, D. U. Apakama, A. Soroush, A. Sakhuja, R. Freeman, C. R. Horowitz, L. D. Richardson, G. N. Nadkarni, E. Klang, Evaluating and addressing demographic disparities in medical large language models: a systematic review, International Journal for Equity in Health 24 (2025).

[6] M. Bucholc, C. James, A. A. Khleifat, A. Badhwar, N. Clarke, A. Dehsarvi, C. R. Madan, S. J. Marzi, C. Shand, B. M. Schilder, S. Tamburin, H. M. Tantiangco, I. Lourida, D. J. Llewellyn, J. M. Ranson, Artificial intelligence for dementia research methods optimization, Alzheimer's & Dementia 19 (2023) 5934–5951.

[7] R. Rawat, H. McBride, D. Nirmal, R. Ghosh, J. Moon, D. Alamuri, S. O'Brien, K. Zhu, Diversitymedqa: Assessing demographic biases in medical diagnosis using large language models, ArXiv abs/2409.01497 (2024).

[8] H. Ahsan, A. S. Sharma, S. Amir, D. Bau, B. C. Wallace, Elucidating mechanisms of demographic bias in llms for healthcare, ArXiv abs/2502.13319 (2025).

[9] S. C. Seong, Y.-Y. Kim, Y.-H. Khang, J. H. Park, H.-J. Kang, H. Lee, C.-H. Do, J.-S. Song, J. H. Bang, S. Ha, E.-J. Lee, S. A. Shin, Data resource profile: The national health information database of the national health insurance service in south korea, International Journal of Epidemiology 46 (2016) 799–800.

[10] H. Kim, H.-W. Chun, S. Kim, B.-Y. Coh, O.-J. Kwon, Y.-H. Moon, Longitudinal study-based dementia prediction for public health, International Journal of Environmental Research and Public Health 14 (2017) 983.

7. Reviewer's comments and responses

1. Clarify Conceptual Framework: Please distinguish between sex (biological) and gender (social construct) in your analysis. Additionally, consider acknowledging that some "gender-specific risk factors" (e.g., differential schizophrenia diagnosis rates) may themselves reflect societal biases rather than inherent biological differences.

- I revised the term "gender" to "sex" in the paper. What I want to deal with is "biological sex".

2. Enhance Transparency: Include the exact prompts used in your experiments. This will improve reproducibility and help others build upon your work.

- I am preparing another journal article with the future work and it should be contain the exact prompts. By the way, I think the section "3.2 LLM-Based Prediction System Architecture and Experimental Setup" contains the detailed explanation about the process.

3. Add Visual Comparisons: Consider adding figures or charts comparing model performances to make the results more accessible and impactful.

- I added a table (Table 3) to explain performance comparison explicitly.

4. Consider Intersectionality: While you mention evaluating across demographic subgroups, consider discussing how intersectional identities (e.g., age × gender) might affect predictions and fairness in future work.

- I updated "Future Plan" to reflect the comments.