

Longitudinal multisource clinical model for early lung cancer risk stratification and screening

Chia-Hui Chien,^{1,2,3} Shih-Chuan Chang,⁴ Yung-Chun Chang ,^{4,5}
Yu-Chuan Li ,^{2,3,6,7}

To cite: Chien C-H, Chang S-C, Chang Y-C, *et al.* Longitudinal multisource clinical model for early lung cancer risk stratification and screening. *BMJ Health Care Inform* 2026;**33**:e101989. doi:10.1136/bmjhci-2025-101989

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2025-101989>).

Received 04 December 2025
Accepted 11 January 2026



© Author(s) (or their employer(s)) 2026. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ Group.

For numbered affiliations see end of article.

Correspondence to
Dr Yung-Chun Chang;
changyc@tmu.edu.tw

Dr Yu-Chuan Li;
jack@tmu.edu.tw

ABSTRACT

Objectives Lung cancer is the leading cause of cancer-related mortality worldwide, with poor prognosis largely due to late-stage diagnosis. Current screening methods such as low-dose CT face accessibility and cost barriers in resource-limited settings. This study develops a lightweight multichannel convolutional neural network for lung cancer screening support through longitudinal risk stratification using routine pre-diagnostic healthcare data.

Methods We conducted a retrospective cohort study using Taiwan's National Health Insurance Research Database, comprising 99 615 individuals (575 lung cancer cases; 99 040 non-cancer controls). Diagnostic codes, medication records and medical orders within a 36-month observation window were extracted. Log-likelihood ratio feature selection was implemented to reduce dimensionality, achieving 99.8% reduction in computational requirements while retaining clinical relevance. A multichannel Convolutional Neural Network (CNN) architecture was designed to process these heterogeneous data modalities simultaneously.

Results The proposed method achieved an F₁-score of 0.5738, precision of 0.7149, Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.8316 and Area Under the Precision-Recall Curve (AUPRC) of 0.1617, outperforming baseline methods in precision and F₁-score. Ablation studies confirmed that medical orders provide primary predictive value, while medication features contribute limited discriminative signal in the pre-diagnostic phase. SHapley Additive exPlanations analysis revealed that routine healthcare utilisation patterns, rather than cancer-specific features, drive risk stratification.

Discussion The lightweight architecture enables deployment in resource-constrained clinical environments while maintaining robust performance, offering potential as a preliminary screening tool to identify high-risk individuals for further diagnostic examination.

Conclusion Efficient deep learning models using routine clinical data can facilitate lung cancer risk stratification and screening, providing a scalable solution for clinical implementation.

INTRODUCTION

Lung cancer remains the leading cause of cancer-related mortality worldwide. In 2022, approximately 2.5 million new cases were diagnosed globally, resulting in

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Deep learning models using electronic health records have shown promise for cancer risk prediction, but computational demands and interpretability limitations hinder deployment in resource-constrained clinical settings where preliminary screening tools are most needed.

WHAT THIS STUDY ADDS

⇒ This study demonstrates that a lightweight multichannel Convolutional Neural Network (CNN) with log-likelihood ratio feature selection can achieve superior precision (0.7149) and F₁-score (0.5738) for lung cancer detection, while reducing computational requirements by 99.8%, and that routine healthcare utilisation patterns rather than cancer-specific features drive pre-diagnostic risk stratification.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The framework offers a scalable approach for integrating artificial intelligence (AI)-driven lung cancer risk assessment into existing electronic health record systems, potentially enabling identification of high-risk individuals who would benefit most from further diagnostic examination such as low-dose CT screening.

1.8 million deaths.¹ The situation is particularly concerning in Taiwan, where lung cancer accounts for approximately 15 000 new diagnoses and over 9000 deaths annually.² The fundamental challenge in lung cancer management lies in the timing of diagnosis. Early-stage detection significantly improves patient prognosis, with survival rates exceeding 80% when tumours are identified before metastatic spread.³ However, the asymptomatic nature of early-stage disease creates a detection paradox where patients typically present with symptoms only after progression to advanced stages. Low-dose CT (LDCT) has shown promise in reducing lung cancer mortality in high-risk populations, with the National Lung Screening Trial

demonstrating approximately 20% mortality reduction compared with chest radiography.^{4,5} Yet widespread LDCT implementation faces substantial barriers including high operational costs, limited accessibility in resource-constrained healthcare systems and false-positive rates reaching 24% in some screening programmes.⁶ These limitations highlight the need for efficient preliminary screening tools that can identify high-risk individuals who would benefit most from further diagnostic examination.

Artificial intelligence has created opportunities to transform medical diagnostics through analysing complex clinical data patterns.^{7,8} Deep learning architectures have achieved remarkable performance in medical applications, while electronic health records and administrative healthcare databases offer rich longitudinal information including diagnostic histories, medication patterns and medical orders.^{9,10} Taiwan's National Health Insurance Research Database (NHIRD) represents one of the world's most comprehensive population-based healthcare datasets, containing longitudinal clinical information for virtually the entire Taiwanese population and providing an exceptional foundation for developing AI-driven prediction models.^{11,12} However, realising the potential of such heterogeneous clinical data requires analytical frameworks capable of processing multiple data modalities, while preserving temporal relationships. Traditional machine learning approaches often struggle with the complexity and dimensionality of longitudinal clinical data. Moreover, computational demands of sophisticated deep learning models may limit their practical deployment in resource-constrained clinical environments, where efficient preliminary screening tools are most needed.

In this study, we focus on longitudinal risk stratification using routine pre-diagnostic healthcare data to support lung cancer screening, rather than direct detection of cancer presence at the time of inference. Our approach integrates 3 years of longitudinal patient data from NHIRD, incorporating diagnostic codes, medication records and medical orders to create comprehensive clinical profiles. A systematic feature selection strategy based on log-likelihood ratio was integrated with the CNN architecture to substantially reduce computational requirements while retaining clinical relevance, enabling deployment in resource-limited settings. Through this integration of efficient deep learning and population-scale healthcare data, we aim to explore whether lightweight CNN architectures can effectively extract temporal patterns from routine clinical records for lung cancer risk stratification, while maintaining the interpretability essential for clinical adoption.

MATERIALS AND METHODS

Study design and population

This retrospective cohort study used Taiwan's NHIRD, encompassing over 99% of Taiwan's 23 million residents.¹¹ The NHIRD contains longitudinal diagnostic

codes, prescription records and medical orders. Data are deidentified and maintained by the Health and Welfare Data Science Center.

The study population was derived from the NHIRD 100 000-participant sampling file (1999–2013). The lung cancer cohort comprised patients with confirmed diagnoses identified (International Classification of Diseases (ICD)-9 codes 162.xx). The index date was the first lung cancer diagnosis, with a 36-month observation window extracted retrospectively. The non-cancer cohort consisted of patients without any cancer diagnoses, with the latest healthcare record as index date.

Inclusion criteria required continuous National Health Insurance enrolment for 3 years prior to index date, age of ≥ 18 years, and at least one healthcare encounter per year. Patients with incomplete coverage, inconsistent codes or insufficient data were excluded. The final population comprised 99 615 individuals (575 lung cancer cases; 99 040 non-cancer controls). Demographic characteristics (table 1) were not direct model inputs but may indirectly influence predictions through healthcare utilisation patterns.

Data preprocessing and feature selection

Clinical data from the 36-month observation window encompassed diagnostic codes, medication prescriptions and medical orders. The 36-month period was selected based on clinical evidence that pre-diagnostic utilisation changes emerge 2–3 years before diagnosis, alignment with routine care patterns and NHIRD data availability.¹³ Medical orders predominantly represent outpatient pre-diagnostic investigations (chest X-ray, laboratory tests, imaging) rather than inpatient procedures. Diagnostic data used ICD-9 codes, medications used Anatomical Therapeutic Chemical codes¹⁴ and medical orders used standardised procedure codes.

Feature selection employed Log-Likelihood Ratio (LLR) to identify differential occurrence patterns between cohorts.¹⁵ The top 10 features per data modality per cohort were selected, yielding 60 features total (10 features $\times$ 3 modalities $\times$ 2 cohorts; online supplemental table 1). This achieved a 99.8% reduction in computational requirements while retaining clinically relevant predictors. Each feature was mapped across 36 monthly intervals to create temporal matrices capturing clinical event evolution.¹⁶

Model architecture

We employ a multichannel CNN architecture, termed Longitudinal Multisource Clinical Model (LUMIC), to jointly model heterogeneous clinical data while preserving temporal structure. The architecture integrates six input channels from LLR-selected features across three data modalities (diagnostic codes, medications, medical orders) for both cohorts, yielding an input tensor of dimensions (batch, 6, 10, 36) representing channels, features and monthly intervals.

Table 1 Demographic characteristics of the study population

Variables	Cohort population (n=99615)		P value
	Lung cancer	Non-cancer	
Patient number	575 (0.58%)	99 040 (99.42%)	
Gender			<0.0001*
Male	332 (0.33%)	48 796 (48.98%)	
Female	243 (0.24%)	50 244 (50.44%)	
Age, mean (SD), years	67 (15.01)	39 (20.39)	<0.0001†
18–34	20 (0.02%)	39 627 (39.78%)	
34–65	226 (0.23%)	47 123 (47.31%)	
>65	329 (0.33%)	12 290 (12.34%)	
Monthly insured salary, mean (SD), NTD	17 280 (18096)	17 280 (21619.274)	0.0068†
0–20 000	308 (0.31%)	56 465 (56.68%)	
20 000–40 000	220 (0.22%)	30 227 (30.34%)	
>40 000	47 (0.05%)	12 348 (12.40%)	

NTD, New Taiwan Dollar (the currency used in Taiwan).
 * χ^2 test.
 †Independent samples t-test.

The convolutional architecture uses two layers optimised for temporal clinical data. The first layer processes 6-channel input with 32 output channels using 3×3 kernels, followed by Rectified Linear Unit (ReLU) activation and 2×2 max-pooling. The second layer expands to 64 channels with identical configurations.¹⁷ After convolution-pooling, feature maps are flattened for dense layer processing. The architecture incorporates dropout and batch normalisation to prevent overfitting. The final layer employs softmax activation for binary classification (figure 1).

Model training and evaluation

Model training employed stratified 10-fold cross-validation maintaining proportional cohort representation. Hyperparameter optimisation used grid search targeting macro-averaged F_1 -score maximisation. Key hyperparameters included learning rate, batch size, dropout rate and epochs, with early stopping to prevent overfitting.

Performance evaluation employed metrics for imbalanced classification: precision, recall, macro-averaged F_1 -score, Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision-Recall Curve (AUPRC), with AUPRC emphasised given 0.58% positive cases. AUROC and AUPRC are threshold-free; threshold-dependent metrics used 0.5 default, adjustable for deployment. Comparative analysis evaluated the framework against traditional machine learning (Naive Bayes, Logistic Regression, Linear SVM, Decision Tree, Random Forest, KNN), DeepFFM for factorisation-based modelling, and ConvBERT for transformer transferability.^{18 19} LUMIC integrates Electronic Medical Record (EMR) input, LLR feature selection and multi-channel CNN as a unified system. Baseline models used the complete original feature set without LLR. Traditional

models aggregated temporal features into summary statistics; DeepFFM and ConvBERT used temporal sequences with embeddings.

Model interpretability used SHAP (SHapley Additive exPlanations) analysis to quantify feature contributions.^{20 21} SHAP values were computed across all six input channels, generating temporal importance matrices (6 channels×10 features×36 time periods) visualising how clinical events influenced classification decisions.

RESULT AND DISCUSSION

Results

The proposed LUMIC was evaluated against traditional machine learning and deep learning baseline models using stratified 10-fold cross-validation. Table 2 presents the comparative performance across all evaluated methods.

Among traditional machine learning approaches, Logistic Regression achieved the highest F_1 -score of 0.5693 with balanced precision-recall trade-off, while Random Forest and Linear SVM demonstrated strong discriminative ability (AUROC: 0.8587 and 0.8400, respectively) but minimal sensitivity. Decision Tree and KNN exhibited weak discriminative performance with AUROC values below 0.55. Among deep learning baselines, DeepFFM achieved the highest recall (0.8413) and AUROC (0.9137) but with modest precision (0.5188), while ConvBERT performed poorly across all metrics (F_1 -score: 0.2029, AUROC: 0.5435), suggesting transformer-based architectures require substantial adaptation for structured longitudinal clinical records.

The proposed LUMIC achieved the highest precision (0.7149) and F_1 -score (0.5738) among all methods, with

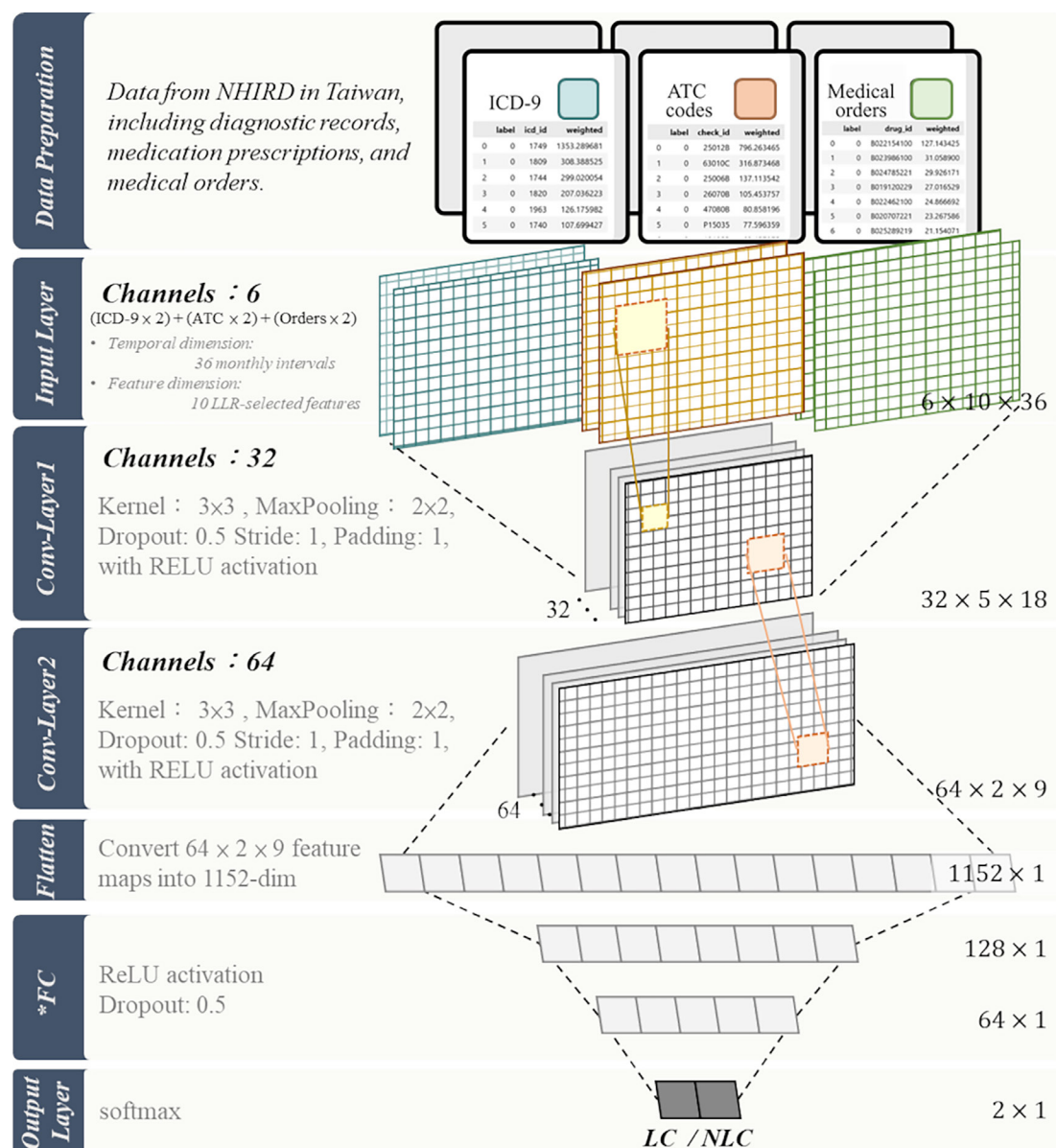


Figure 1 Proposed Longitudinal Multisource Clinical Model (LUMIC) architecture for early cancer detection. ATC, Anatomical Therapeutic Chemical; FC, fully connected layers; LC, lung cancer; NHIRD, National Health Insurance Research Database; NLC, non-lung cancer.

AUPRC of 0.1617 representing optimal performance for imbalanced classification. While DeepFFM achieved higher recall and AUROC, its lower precision would generate more false positive referrals in clinical deployment. In terms of computational efficiency, the runtime per cross-validation fold for LUMIC (54.57s) was comparable to that of DeepFFM (47.15s) and substantially faster than that of ConvBERT (756.77s) and most traditional approaches.

Ablation analysis examined each data modality's contribution, as shown in table 3. Medical order omission caused the most substantial degradation, with F1-score decreasing to 0.5345 and AUPRC dropping to 0.1045, confirming procedural data as essential for capturing clinical evaluation pathways preceding formal diagnosis. ICD omission resulted in a moderate reduction (F1-score:

0.5452). Interestingly, medication omission yielded marginally higher F1-score (0.5775) and AUPRC (0.1670) compared with the full model. This finding suggests that pre-diagnostic medication patterns may share similar characteristics across various conditions, including symptomatic treatments common to early-stage respiratory complaints and other potential malignancies, thereby introducing noise rather than discriminative signal for lung cancer-specific detection. Diagnostic codes and medical orders thus carry primary predictive value in the pre-diagnostic phase, while medication features may become more discriminative only after cancer-specific pharmacological interventions commence.

SHAP analysis quantified temporal feature contributions (figure 2). Non-cancer cohort features (left column) demonstrated substantially greater contributions than

Table 2 Performance comparison of LUMIC and baseline models

Method	Precision	Recall	F ₁ -score	AUROC	AUPRC	Runtime (s)
Traditional ML						
Naive Bayes	0.5387	0.6337	0.5577	0.6960	0.0487	0.55
Logistic Regression	0.5450	0.6912	0.5693	0.8029	0.1566	217.04
Linear SVM	0.4971	0.5000	0.4986	0.8400	0.0742	151.42
Random Forest	0.4971	0.5000	0.4986	0.8587	0.0409	158.62
Decision Tree	0.5146	0.5162	0.5153	0.5263	0.0082	303.58
KNN	0.4971	0.5000	0.4985	0.5442	0.0114	135.89
Deep learning						
DeepFFM	0.5188	0.8413	0.5040	0.9137	0.1015	47.15
ConvBERT	0.2006	0.5000	0.2029	0.5435	0.0081	756.77
LUMIC	0.7149	0.5461	0.5738	0.8316	0.1617	54.57

Bold typeface signifies best performance in each column.
AUPRC, Area Under the Precision-Recall Curve; AUROC, Area Under the Receiver Operating Characteristic Curve; LUMIC, Longitudinal Multisource Clinical Model.

cancer cohort features (right column), which showed sparse activity across all three modalities. Among non-cancer features, the examination channel exhibited the strongest signal density, while both ICD and medication channels displayed heterogeneous patterns with mixed positive and negative contributions throughout the observation window. The sparse contributions from cancer cohort features indicate that cancer-specific clinical patterns emerge only proximal to formal diagnosis, supporting the framework's rationale of leveraging routine healthcare engagement for early risk stratification.

DISCUSSION

The proposed LUMIC demonstrates that lightweight deep learning architectures can effectively leverage routine healthcare data for lung cancer risk stratification. By achieving the highest precision (0.7149) and F₁-score (0.5738) among evaluated methods while maintaining competitive inference time, this approach offers a practical balance between predictive performance and clinical deployability. While DeepFFM achieved superior recall (0.8413) and AUROC (0.9137), its lower precision (0.5188) would generate approximately 37% more false positive referrals. In cancer screening contexts,

false positives lead to unnecessary diagnostic procedures, patient anxiety and healthcare resource consumption.²² LUMIC prioritises precision while maintaining reasonable sensitivity, reducing follow-up burden for patients without disease. Although LUMIC outperformed baseline models, its moderate F₁-score (0.5738) with higher precision (0.7149) than recall (0.4733) reflects deliberate prioritisation of precision appropriate for population-level screening. This performance profile would be insufficient for standalone diagnostic decision-making, which requires higher sensitivity to minimise false negatives. These trade-offs reflect fundamental differences between screening contexts, where moderate false positive rates are acceptable, and diagnostic contexts, where missed cancers carry severe consequences. The substantial baseline differences in table 1 (age, sex, insured salary) represent established risk factors whose effects may be partially captured through healthcare utilisation patterns, enabling risk identification beyond demographic factors alone.

ConvBERT's poor performance (F₁-score: 0.2029, AUROC: 0.5435) demonstrates that transformer architectures designed for natural language processing require substantial adaptation for structured longitudinal clinical

Table 3 Performance impact of individual data modality omission

	Precision	Recall	F ₁ -score	AUROC	AUPRC	Runtime (s)
LUMIC	0.7149	0.5461	0.5738	0.8316	0.1617	54.57
ICD omission	0.6907	0.5270	0.5452	0.8333	0.1360	38.84
Medication omission	0.7312	0.5479	0.5775	0.8280	0.1670	39.35
Medical order omission	0.6930	0.5209	0.5345	0.8077	0.1045	39.84

Bold typeface signifies best performance in each column.
AUPRC, Area Under the Precision-Recall Curve; AUROC, Area Under the Receiver Operating Characteristic Curve; ICD, International Classification of Diseases; LUMIC, Longitudinal Multisource Clinical Model.

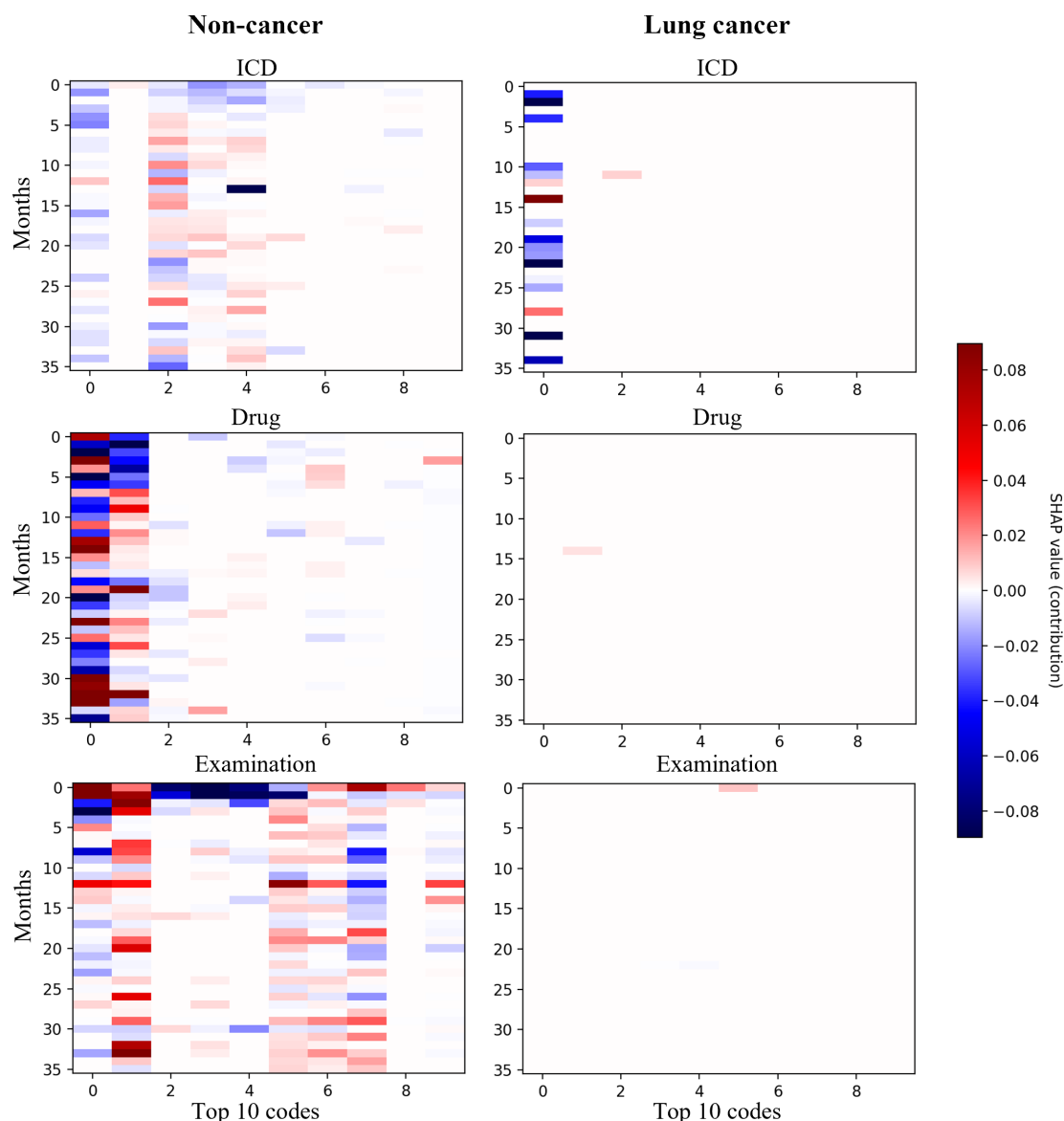


Figure 2 Temporal SHAP contributions comparing non-cancer and lung cancer cohort features. Red indicates positive contributions towards lung cancer classification; blue indicates negative contributions. ICD, International Classification of Diseases; SHAP, SHapley Additive exPlanations.

records.^{18 19} Its inferior performance reflects architectural limitations including absence of explicit temporal alignment and reliance on dense token representations incompatible with sparsely coded, event-based clinical records. These findings support CNN-based approaches for modelling localised temporal patterns in administrative healthcare data.

The ablation study provides clinically interpretable insights into data modality contribution. Medical order omission caused substantial performance degradation (F1-score: 0.5345, AUPRC: 0.1045), confirming procedural data capture critical information about clinical evaluation pathways preceding diagnosis. In contrast, medication omission marginally improved performance (F1-score: 0.5775 vs 0.5738), suggesting pre-diagnostic medication patterns may introduce noise rather than discriminative signal. Symptomatic treatments and medications common

to multiple conditions likely share similar characteristics across patient populations, limiting lung cancer specificity prior to definitive diagnosis. The marginal improvement after removing medication features likely reflects noise reduction under LLR-based top-K selection, where heterogeneous pre-diagnostic medication patterns may be under-represented when only the top-ranked features are retained. These findings suggest selectively including clinically informative medication categories may be preferable for lightweight screening models.

SHAP analysis revealed distinct temporal patterns validating the model's clinical reasoning. The identified top-ranked predictors across diagnostic codes, medical orders and medications provide insight into early lung cancer risk assessment. High-weight features reflect increased diagnostic testing, respiratory-related encounters and intensified healthcare utilisation, rather

than cancer-specific confirmation, consistent with prior EHR-based cancer prediction studies showing utilisation trajectory changes precede formal diagnosis and capture pre-diagnostic evaluation processes. The framework is intended to operate upstream of LDCT during routine healthcare encounters, performing longitudinal risk stratification to identify individuals at elevated future risk who may benefit from prioritised LDCT referral. By functioning as a screening support rather than a diagnostic tool, the model aims to improve screening programme efficiency by focusing diagnostic resources on individuals most likely to benefit. Non-cancer cohort features demonstrated substantially greater contributions than cancer cohort features, with the examination channel exhibiting strongest temporal dynamics. This indicates the model primarily leverages routine healthcare utilisation patterns (chest radiography, laboratory tests, outpatient consultations) rather than cancer-specific features. The sparse contributions from cancer cohort features confirm cancer-specific patterns emerge only proximal to diagnosis, supporting the framework's design rationale of capturing pre-diagnostic healthcare engagement.

LLR-based feature selection achieves critical computational efficiency for clinical implementation. The 99.8% reduction in memory requirements (from 1163.44 GB to 2.07 GB) enables deployment in resource-constrained environments where sophisticated deep learning models would be impractical.²³ Combined with competitive runtime (54.57s), this positions the framework for EHR system integration supporting real-time risk assessment during routine encounters. Our findings have implications for lung cancer screening programmes. Current LDCT screening faces implementation challenges including high costs, limited accessibility and 24% false positive rates.⁶ LUMIC offers a complementary strategy for identifying high-risk individuals who would benefit most from imaging studies, potentially optimising resource allocation.²⁴

Several limitations warrant consideration. The reliance on NHIRD data through 2013 may not reflect recent advances in lung cancer diagnosis and treatment, and ICD-9 coding has since transitioned to ICD-10. This study uses ICD-9 because the NHIRD dataset spans 1999–2013. Notably, the majority of high-weight diagnostic features—including malignant neoplasms, chronic respiratory diseases and systemic comorbidities—map directly to established ICD-10 codes, demonstrating feature representation generalisability to modern coding systems. Nevertheless, evolving clinical practices, diagnostic workflows and coding standards may introduce concept drift, potentially affecting model calibration. This study should be viewed as methodological proof-of-concept demonstrating feasibility of longitudinal, multisource modelling using administrative healthcare data, not intended for immediate deployment without local retraining. Future research should prioritise validation on contemporary ICD-10 datasets to ensure temporal robustness and clinical utility.

The absence of lifestyle and environmental risk factors, including smoking history and occupational exposures, limits incorporation of established risk factors substantially influencing lung cancer development.²⁵ The geographic specificity of our Taiwanese population raises generalisability questions regarding populations with different demographics, healthcare structures and disease epidemiology.²⁶ External validation using diverse international datasets will be essential for establishing broader applicability. The substantial class imbalance (0.58% positive cases) presents inherent challenges for model training and evaluation.

Despite these limitations, this study establishes a foundation for leveraging routine clinical data in cancer risk prediction. The combination of robust predictive performance, computational efficiency and clinical interpretability positions LUMIC as a valuable complement to existing screening strategies. Future research should focus on external validation, incorporating additional risk factors and prospective clinical utility evaluation.²⁷

CONCLUSION

This study demonstrates that lightweight deep learning architectures can effectively leverage routine clinical data for early lung cancer detection. The proposed LUMIC achieved the highest precision (0.7149) and F₁-score (0.5738) among evaluated methods, with AUPRC of 0.1617 representing optimal performance for imbalanced classification. The multimodal approach successfully integrated 3 years of diagnostic codes, medication records and medical order information, with ablation studies confirming that procedural data provide primary predictive value while pre-diagnostic medication patterns contribute limited discriminative signal.

The SHAP interpretability analysis validated the model's clinical reasoning by revealing that routine healthcare utilisation patterns, rather than cancer-specific features, drive risk stratification in the pre-diagnostic phase. This finding supports the framework's potential as a preliminary screening tool that identifies high-risk individuals through existing healthcare engagement patterns. The computational efficiency achieved through systematic feature selection, reducing memory requirements by 99.8%, enables deployment in resource-constrained clinical environments, where sophisticated deep learning models would otherwise be impractical.

Future implementation will require external validation using diverse international datasets and incorporation of additional risk factors including lifestyle and environmental exposures. Despite these considerations, the demonstrated feasibility of leveraging routine healthcare data for lung cancer risk prediction represents a step towards more accessible and cost-effective approaches to early cancer detection.

Author affiliations

¹Department of Computer Science, Middlesex University, London, UK

²International Center for Health Information Technology, Taipei Medical University, Taipei City, Taiwan

³Graduate Institute of Biomedical Informatics, College of Medical Science & Technology, Taipei Medical University, Taipei City, Taiwan

⁴Graduate Institute of Data Science, Taipei Medical University, Taipei City, Taiwan

⁵Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei City, Taiwan

⁶Department of Dermatology, Wan Fang Hospital, Taipei City, Taiwan

⁷Research Center for Artificial Intelligence in Medicine, Taipei Medical University, Taipei City, Taiwan

Contributors C-HC: writing—original draft, validation, investigation, formal analysis, data curation, methodology. S-CC: validation, formal analysis, methodology. Y-CC: writing—original draft, review and editing, resources, validation, investigation, formal analysis, conceptualisation, supervision, project administration. Y-CL: writing—original draft, review and editing, resources, conceptualisation, methodology, project administration. Y-CC is the guarantor.

Funding This research was supported by the National Science and Technology Council (NSTC) of Taiwan under Grant Nos. NSTC 113-2221-E-038-019-MY3, NSTC 114-2410-H-038-034-MY3 and NSTC 114-2622-8-038-002-TD1. Additional support was provided by Wan Fang Hospital under Grant No. 113-wf-eva-15 and by the Ministry of Education (MOE) of Taiwan.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was approved by the Institutional Review Board of Taipei Medical University (TMU-JIRB No.: N202005106), with informed consent waived due to full de-identification of all data.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The datasets generated and/or analysed during this study are available from the corresponding author upon reasonable request.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <https://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Yung-Chun Chang <https://orcid.org/0000-0002-9634-8380>

Yu-Chuan Li <https://orcid.org/0000-0001-6497-4232>

REFERENCES

- Bray F, Laversanne M, Sung H, *et al*. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024;74:229–63.
- Welfare, T.M.o.H.a. Cancer registry annual report. 2021. Available: <https://www.mohw.gov.tw/>
- Nam B, Hamm D, Katurakes N, *et al*. Lung Cancer Screening: Early Detection Decreases Mortality. *Dela J Public Health* 2024;10:22–4.
- Jonas DE, Reuland DS, Reddy SM, *et al*. Screening for Lung Cancer With Low-Dose Computed Tomography: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA* 2021;325:971–87.
- The National Lung Screening Trial Research Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. *N Engl J Med* 2011;365:395–409.
- Oudkerk M, Liu S, Heuvelmans MA, *et al*. Lung cancer LDCT screening and mortality reduction - evidence, pitfalls and future perspectives. *Nat Rev Clin Oncol* 2021;18:135–51.
- Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature New Biol* 2017;542:115–8.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.
- Litjens G, Kooi T, Bejnordi BE, *et al*. A survey on deep learning in medical image analysis. *Med Image Anal* 2017;42:60–88.
- Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc* 2018;25:1419–28.
- Lin L-Y, Warren-Gash C, Smeeth L, *et al*. Data resource profile: the National Health Insurance Research Database (NHIRD). *Epidemiol Health* 2018;40:e2018062.
- Hsieh C-Y, Su C-C, Shao S-C, *et al*. Taiwan's National Health Insurance Research Database: past and future. *Clin Epidemiol* 2019;11:349–58.
- White B, Renzi C, Rafiq M, *et al*. Does changing healthcare use signal opportunities for earlier detection of cancer? A review of studies using information from electronic patient records. *Cancer Epidemiol* 2022;76:102072.
- Organization, W.H. WHO collaborating centre for drug statistics methodology, ATC/DDD index 2023. 2023.
- Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. in *icml*. Citeseer; 1997.
- Javid H, Mariam A, Khademi G, *et al*. Identification of robust deep neural network models of longitudinal clinical measurements. *NPJ Digit Med* 2022;5:106.
- Catling FJR, Wolff AH. Temporal convolutional networks allow early prediction of events in critical care. *J Am Med Inform Assoc* 2020;27:355–65.
- Guo H, Tang R, Ye Y, *et al*. DeepFM: a factorization-machine based neural network for CTR prediction. Proceedings of the twenty-sixth International Joint Conference on Artificial Intelligence (IJCAI); 2017.
- Jiang Z-H, *et al*. Convbert: Improving bert with span-based dynamic convolution. *Adv Neural Inf Process Syst* 2020;33:12837–48.
- Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: *Advances in neural information processing systems*, 30. 2017.
- Alkhanbouli R, Matar Abdulla Almadhaani H, Alhosani F, *et al*. The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions. *BMC Med Inform Decis Mak* 2025;25:110.
- Dabbous FM, Dolecek TA, Berbaum ML, *et al*. Impact of a False-Positive Screening Mammogram on Subsequent Screening Behavior and Stage at Breast Cancer Diagnosis. *Cancer Epidemiol Biomarkers Prev* 2017;26:397–403.
- Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* 2017;376:2507–9.
- Chandran U, Reps J, Yang R, *et al*. Machine Learning and Real-World Data to Predict Lung Cancer Risk in Routine Care. *Cancer Epidemiol Biomarkers Prev* 2023;32:337–43.
- Malhotra J, Malvezzi M, Negri E, *et al*. Risk factors for lung cancer worldwide. *Eur Respir J* 2016;48:889–902.
- Torre LA, Bray F, Siegel RL, *et al*. Global cancer statistics, 2012. *CA A Cancer J Clinicians* 2015;65:87–108.
- Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019;380:1347–58.