

An exploratory study of multi-channel CNN for early detection of lung cancer from longitudinal healthcare records

Received: 15 September 2025

Accepted: 4 May 2026

Published online: 11 May 2026

Cite this article as: Chien C., Chang S., Chang Y. *et al.* An exploratory study of multi-channel CNN for early detection of lung cancer from longitudinal healthcare records. *Sci Rep* (2026). <https://doi.org/10.1038/s41598-026-52173-8>

Chia-Hui Chien, Shih-Chuan Chang, Yung-Chun Chang & Yu-Chuan Li

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

An Exploratory Study of Multi-Channel CNN for Early Detection of Lung Cancer from Longitudinal Healthcare Records

Chia-Hui Chien^{1,2,3}, Shih-Chuan Chang⁴, Yung-Chun Chang^{4,5,z,2*}, and Yu-Chuan Li^{2,3y,3*}

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

²Graduate Institute of Biomedical Informatics, Taipei Medical University, Taiwan

³International Center for Health Information Technology, Taipei Medical University, Taiwan

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

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

^{z,2} Corresponding author. Fax: +886-2-6620-2589 ext. 16105 (Y.C. Chang).

E-mail address: changyc@tmu.edu.tw (Y.C. Chang).

^{y,3} Corresponding author. Fax: +886-2-2736-1661 ext. 8627 (Y.C. Li).

E-mail address: jack@tmu.edu.tw (Y.C. Li).

ABSTRACT

Lung cancer remains the leading cause of cancer-related mortality worldwide, with poor survival rates due to late-stage detection. Current low-dose computed tomography screening faces barriers including high costs and false-positive rates reaching 24%, while artificial intelligence offers opportunities to enhance early detection through longitudinal clinical data analysis. This study developed a Multi-Channel Convolutional Neural Network (MCNN) for lung cancer risk prediction using Taiwan's National Health Insurance Research Database, encompassing 523,539 patients (2,809 lung cancer, 23,783 other cancer, and 496,947 non-cancer). The MCNN was designed as a lightweight model processing nine channels of diagnostic codes, medications, and medical orders over a three-year observation period. Systematic feature selection reduced estimated feature storage requirements by 99.8%, from approximately 1,184 GB for the full ICD feature space to approximately 2.11 GB for the selected features, while retaining clinical relevance. Model performance was assessed using stratified 10-fold cross-validation against seven machine learning baselines, and interpretability was examined through SHAP analysis. The MCNN achieved an F₁-score of 66.91%, precision of 84.47%, and recall of 59.79%. Ablation studies confirmed multi-modal integration benefits, with diagnostic codes providing primary predictive power. SHAP analysis revealed distinct temporal patterns validating the model's ability to identify pre-diagnostic phases through healthcare engagement patterns. Findings are based on internal validation within a single national database, and key risk factors such as smoking history are not captured in administrative claims data; future evaluation in independent external cohorts is therefore warranted to confirm these findings. The model's high precision minimizes false-positive rates while its computational efficiency and clinical interpretability support practical implementation as a complementary claims-based screening support tool for early cancer detection.

Keywords: Lung Cancer Detection, Multi-channel CNN, Early Screening,

INTRODUCTION

Lung cancer remains the most devastating malignancy worldwide and the leading cause of cancer-related mortality. In 2022, approximately 2.5 million new cases were diagnosed globally, resulting in 1.8 million deaths [1]. In Taiwan, lung cancer accounts for approximately 15,000 new diagnoses and over 9,000 deaths annually, making it a critical healthcare priority [2]. Despite advances in therapeutic interventions, the five-year survival rate remains low due to late-stage detection, whereas early-stage identification yields survival rates exceeding 80% before metastatic spread [3]. The asymptomatic nature of early-stage disease means patients typically present only after cancer has progressed to advanced stages. Low-dose computed tomography (LDCT) has shown promise in reducing lung cancer mortality, with the National Lung Screening Trial demonstrating approximately 20% mortality reduction compared to chest radiography [4, 5]. However, widespread LDCT implementation faces substantial barriers including high costs, limited accessibility, and false-positive rates reaching 24%, which can lead to unnecessary invasive procedures and significant patient distress [6].

Artificial intelligence has created unprecedented opportunities to transform medical diagnostics and risk assessment. Machine learning algorithms have demonstrated remarkable capabilities in analyzing complex medical data patterns [7-10], and deep learning architectures, particularly convolutional neural networks, have achieved breakthrough performance in medical image analysis [11], with clinical-grade examples including diabetic retinopathy screening, skin cancer detection, and pneumonia identification from chest radiographs [12]. Beyond imaging, AI has shown considerable promise in analyzing electronic health records (EHRs) and administrative healthcare databases [13, 14], which contain rich longitudinal information including diagnostic histories, medication usage, and medical order data [15]. Integration of such multidimensional data through machine learning offers the potential to identify subtle early warning signals preceding disease manifestation [16, 17]. Taiwan's National Health Insurance Research Database (NHIRD) represents one of the world's most extensive population-based healthcare datasets, covering virtually the entire Taiwanese population [18] and capturing detailed patient journeys across multiple healthcare encounters, creating unique opportunities for population-scale research [19].

However, realizing the full potential of heterogeneous clinical data requires analytical frameworks capable of simultaneously processing multiple data modalities while preserving temporal relationships. Traditional machine learning approaches often struggle with the complexity and dimensionality of longitudinal clinical data when integrating diagnostic codes, medication histories, and clinical procedure codes into unified predictive models. This analytical challenge necessitates the development of novel deep

learning architectures specifically designed to handle multi-modal temporal data and extract meaningful patterns from years of accumulated clinical information.

The present study addresses these challenges by developing a novel Multi-Channel Convolutional Neural Network (MCNN) architecture specifically designed to integrate multidimensional clinical data for early lung cancer risk prediction. While recurrent or attention-based architectures can capture temporal dependencies, they involve greater computational complexity and are less suited for large administrative healthcare datasets. The proposed MCNN leverages three years of longitudinal NHIRD data incorporating diagnostic codes, medication records, and medical orders, and is designed to efficiently capture temporal patterns through structured feature channels while supporting computational efficiency and model interpretability. Because administrative claims data do not directly capture key risk factors such as smoking history or occupational exposure, the model is intended as a complementary claims-based screening support tool rather than a replacement for clinical risk models that incorporate these variables. We hypothesize that this multi-channel convolutional framework can achieve competitive predictive performance while maintaining a substantially lower computational footprint, advancing the accessibility and efficiency of lung cancer screening programs.

MATERIALS AND METHODS

Study Population

This study used data from the NHIRD, a comprehensive population-based healthcare claims database encompassing over 99% of Taiwan's 23 million residents [18]. The NHIRD contains longitudinal clinical information including diagnostic, procedural, and prescription data from 1995 to present. All data are de-identified and maintained by the Health and Welfare Data Science Center of Taiwan's Ministry of Health and Welfare [19]. This study was approved by the Institutional Review Board of Taipei Medical University (TMU-JIRB No.: N202005106); informed consent was waived as all NHIRD data were fully de-identified prior to analysis. The study population was derived from the NHIRD Clinical Declaration and In-Hospital Declaration, consisting of one million files covering January 1, 1999 to December 31, 2013, with a three-year retrospective observation window established for each patient. Patients were classified into three cohorts based on their diagnostic history.

The lung cancer cohort comprised patients with confirmed lung cancer diagnoses identified through the International Classification of Diseases, Ninth Revision (ICD-9) codes 162.xx, as Taiwan's health insurance database transitioned to ICD-10 only in 2016. Inclusion required at least one inpatient hospitalization record with the corresponding ICD-9 code; outpatient-only

records were insufficient to ensure diagnostic specificity, as inpatient coding is generally considered more reliable for cancer case identification in administrative claims data. Prior validation against the Taiwan National Cancer Registry reported a positive predictive value of 95% for lung cancer diagnosis in NHIRD, supporting this ascertainment approach [20]. The other cancer cohort included patients diagnosed with malignancies other than lung cancer, serving as a control group to distinguish lung cancer-specific patterns from general cancer-related healthcare engagement. The non-cancer cohort consisted of patients without any cancer diagnoses, representing the general population baseline. Inclusion criteria required continuous NHI enrollment for at least three years prior to the index date, age ≥ 18 years at the start of the observation period, complete demographic information, and at least one healthcare encounter per year during the observation window to ensure adequate data density. Patients were excluded for incomplete insurance coverage, inconsistent or contradictory diagnostic codes, or primary cancers diagnosed or treated outside the NHI system.

The three-year observation window was anchored differently per cohort to ensure comparable retrospective data capture. For lung cancer and other cancer patients, the 36-month window was calculated retrospectively from the date of first cancer-related hospitalization. For non-cancer patients, it was calculated backward from the last hospitalization date, ensuring no cancer diagnosis occurred throughout the observation period. The index hospitalization was not included as model input; only clinical events within the defined 36-month window prior to the index date were used for feature construction. The final study population is demonstrated as Table 1, which consisted of 523,539 patients, with 2,809 lung cancer patients, 23,783 patients with other cancers, and 496,947 non-cancer patients. Demographic differences across cohorts were statistically significant ($p < 0.0001$ for both gender and age). Lung cancer patients were predominantly male (61.02%) with a mean age of 67.27 years (standard deviation 13.46); the other cancer cohort was nearly balanced by gender with a mean age of 58.91 years (standard deviation 16.12); and the non-cancer cohort was majorly females (54.38%), with a mean age of 37.31 years (standard deviation 22.92).

Table 1. Demographic characteristics of NHIRD study cohorts

Cohort population (n = 523,539)				P value
Variables	Lung cancer	Other cancer	Non cancer	
Patient number	2,809	23,783	496,947	
Gender, n (%)				$p < 0.0001$
Male	1,714 (61.02)	11,961 (50.29)	226,689 (45.62)	
Female	1,095 (38.98)	11,822 (49.71)	270,258 (54.38)	
Age, mean (sd), yr	67.27 (13.46)	58.91(16.1 2)	37.31 (22.92)	$p < 0.0001$

We acknowledge that anchoring the non-cancer observation window to the last hospitalization may introduce selection bias by excluding individuals without hospitalization during the study period. However, this design was necessary to ensure adequate longitudinal records and consistent temporal alignment across cohorts, as individuals without healthcare encounters would not generate sufficient signals for longitudinal analysis. This design may slightly limit generalizability to extremely healthy individuals with minimal healthcare utilization.

Data Preprocessing and Feature Selection

Clinical data extracted from NHIRD were systematically preprocessed to transform heterogeneous healthcare records into structured inputs for deep learning analysis. Initial cleaning addressed missing values, duplicate records, and inconsistent formatting across the 15-year database span. Diagnostic information was standardized using ICD-9 codes, medication data using Anatomical Therapeutic Chemical (ATC) classification codes to ensure consistent pharmaceutical categorization across brand names and formulations [21], and medical order information using standardized procedure codes as recorded in NHIRD. All clinical events within each patient's three-year observation window were chronologically ordered and indexed relative to the study endpoint.

Given the large-scale nature of NHIRD, reducing computational complexity was a primary design priority to enable deployment across diverse clinical settings. Feature selection addressed the high dimensionality of longitudinal clinical data, where thousands of diagnostic, medication, and procedure codes could potentially be extracted. Extremely rare codes were excluded prior statistical ranking to reduce sparsity and improve model stability. The Log-Likelihood Ratio (LLR) method was selected as the primary feature ranking criterion due to its effectiveness in identifying features with differential occurrence patterns between lung cancer patients and control groups [22, 23]. For each potential feature f , the LLR was computed as:

$$\text{LLR}(f) = 2 \times \left[N(f \cap \text{LC}) \log \frac{p(f|\text{LC})}{p(f)} + N(f \cap \neg\text{LC}) \log \frac{p(f|\neg\text{LC})}{p(f)} \right] \quad (1)$$

where $\square$ denotes lung cancer patients, $N(f \cap \text{LC})$ the number of lung cancer cases containing feature f , and $N(f \cap \neg\text{LC})$ its occurrence in non-cancer patients. The conditional probabilities $p(f|\text{LC})$ and $p(f|\neg\text{LC})$ were estimated using maximum likelihood estimation, and $p(f)$ represented overall feature prevalence across the study population. Feature selection was conducted independently for each data modality, with the top 10 features per modality selected by LLR score, yielding 30 primary predictors. This count balanced predictive performance, computational efficiency, and clinical interpretability: retaining the full feature space would substantially increase model dimensionality and computational burden, while equal selection across

modalities prevented dominance from any single data source and facilitated modality-level interpretability in subsequent SHAP analysis. Selected diagnostic features included respiratory-related conditions, oncological diagnoses, and comorbidities associated with lung cancer; medication features encompassed therapeutic classes for respiratory conditions, cancer treatments, and supportive care; and medical order features included tissue biopsy, antigen detection, and tumor excision procedures reflecting early diagnostic workup. Each selected feature was mapped across 36 monthly intervals to create feature matrices capturing both the presence and temporal evolution of clinical events [24], enabling identification of progressive patterns such as increasing respiratory diagnoses or escalating medication usage that may indicate developing pathological processes. The resulting preprocessed dataset consisted of structured feature matrices per patient, optimized for convolutional neural network processing while preserving the temporal and clinical information essential for lung cancer risk prediction.

Multi-Channel Convolutional Neural Network Architecture

The proposed MCNN adopts a unified multi-channel framework that simultaneously processes diagnostic, medication, and medical order data streams while preserving their temporal structure and modality-specific information. The overall architecture was partly inspired by prior studies that applied convolutional neural networks to longitudinal electronic medical record data by transforming sequential healthcare events into image-like matrices, allowing convolutional filters to capture temporal patterns across time [25, 26]. In this study, a relatively compact architecture of two convolutional layers with pooling operations was adopted to balance predictive performance, computational efficiency, and model interpretability when applied to large-scale administrative healthcare databases.

The architecture simultaneously processes nine distinct channels representing the top 10 features for each data modality across three outcome classes: (1) ICD diagnostic codes for lung cancer, other cancer, and non-cancer patients; (2) ATC medication codes for the same three categories; and (3) medical order codes for each patient group. The input tensor has dimensions of (batch, 9, 10, 36), where 9 represents the feature channels, 10 the top-ranked features within each channel, and 36 the monthly intervals across the three-year observation period. The first convolutional layer (Conv1) processes the 9-channel input with 32 output channels using 3×3 kernels, followed by ReLU activation and 2×2 max-pooling. The second convolutional layer (Conv2) expands to 64 channels with the same kernel configuration, again followed by ReLU activation and 2×2 max-pooling. This hierarchical extraction captures progressively complex temporal patterns, from local month-to-month variations to broader disease progression patterns [27]. The resulting feature maps (64, 2, 9) are flattened into a 1,152-dimensional vector for dense layer processing.

The flattened vector encapsulates cross-modal relationships across all nine input channels, enabling fully connected layers to discover interactions between clinical data types and temporal patterns. Dropout layers are placed after convolutional and fully connected layers to reduce overfitting, and batch normalization is applied to stabilize training dynamics. The final classification layer employs softmax activation to produce probability distributions across the three outcome classes, enabling the model to distinguish lung cancer cases from both other cancer and non-cancer patients for clinically relevant risk stratification. The complete architecture is illustrated in Figure 1. MCNN was implemented in Python 3.11.5 using PyTorch 2.8.0+cu128 with a fixed random seed of 42. All experiments were conducted on an NVIDIA RTX 3090 GPU (24 GB).

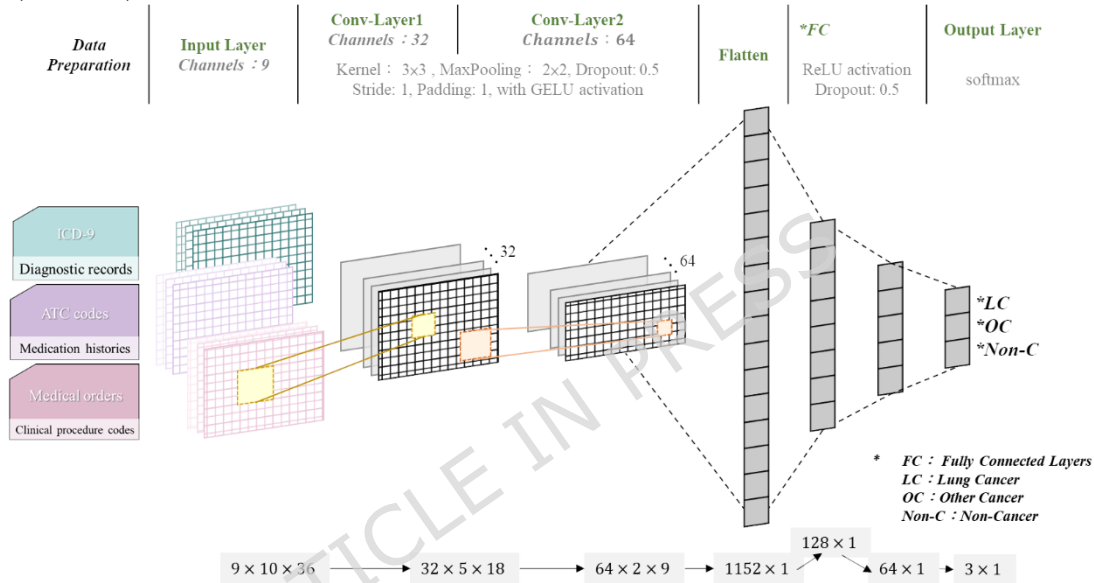


Figure 1. Proposed Multi-Channel CNN Architecture

Systemic Early Lung Cancer Detection Analysis

Model training and validation employed a stratified 10-fold cross-validation approach to ensure robust performance estimation while maintaining proportional representation of all three patient cohorts across training and testing sets. Hyperparameter optimization was conducted using grid search targeting maximization of the F_1 -score. Key hyperparameters included learning rate, batch size, number of convolutional filters, dropout rate, and training epochs. Model training used the Adam optimizer with a learning rate of $1e-4$ and a batch size of 32, with cross-entropy loss as the optimization objective. Training ran for up to 20 epochs with early stopping based on the macro F_1 -score on the validation set with a patience of three epochs. The tuning process was performed prior to the final cross-validation evaluation to identify stable hyperparameter configurations rather than to optimize performance for any specific fold. No explicit resampling strategy was applied during training, as techniques such as SMOTE may disrupt the natural temporal structure of patient records and introduce artificial

healthcare utilization patterns, which would compromise the clinical plausibility of the modeled histories.

Performance evaluation used multiple complementary metrics. Precision measured the proportion of correctly identified lung cancer cases among all positive predictions. Recall quantified sensitivity in detecting actual lung cancer cases. The F_1 -score served as the primary indicator balancing precision and recall, particularly important given the imbalanced dataset. Area under the receiver operating characteristic curve (AUROC) and AUPRC assessed discriminative ability across classification thresholds, with AUPRC especially relevant for imbalanced datasets. Baseline methods included Naive Bayes, Decision Tree, Random Forest, Light Gradient Boosting Machine (LightGBM), DeepFFM, and RETAIN, representing probabilistic, tree-based, ensemble, and specialized deep learning architectures [28, 29]. All baseline models underwent identical preprocessing and feature selection protocols, with hyperparameter tuning using the same grid search methodology. Statistical significance of performance differences between MCNN and baseline models was evaluated using the Wilcoxon signed-rank test across cross-validation folds.

Model Interpretability Analysis

SHAP (SHapley Additive exPlanations) analysis was implemented for post-hoc model explanation, quantifying each feature's contribution to individual predictions by computing its marginal contribution across all possible feature combinations [30]. SHAP is particularly well-suited for healthcare applications where model transparency is essential for clinical acceptance [31]. Analysis was conducted on trained MCNN models following cross-validation, using the Python SHAP library optimized for deep neural networks. SHAP values were computed across all nine input channels, generating individual feature importance matrices (9 channels $\times$ 10 features $\times$ 36 time periods) to visualize how clinical events and their temporal patterns influenced classification decisions. Representative cases from each cohort were selected based on clear diagnostic outcomes and comprehensive clinical data throughout the observation window. SHAP visualizations present temporal heatmaps where positive contributions (red) indicate features increasing lung cancer classification probability and negative contributions (blue) indicate features decreasing it, with color intensity reflecting contribution magnitude.

RESULT AND DISCUSSION

Results

The comprehensive evaluation demonstrated superior MCNN performance across multiple metrics as illustrated in Table 2. Model performance metrics are reported with mean $\pm$ standard deviation with 95% confidence intervals

across the cross-validation folds. The MCNN achieved the highest F_1 -score of 0.6691, precision of 0.8447, and recall of 0.5979. These results represent a substantial improvement over traditional machine learning approaches and competitive performance against the strongest deep learning baseline, reflecting the model's balanced ability to identify true positive lung cancer cases while maintaining high precision in predictions.

Table 2. Performance Comparison of Lung Cancer Detection Models

Method	Precision	Recall	F_1 -score	AUROC	AUPRC
Naive Bayes	0.4307±0.0007 ** (0.4302-0.4312)	0.7438±0.0075 ** (0.7384-0.7491)	0.3124±0.0018 ** (0.3111-0.3136)	0.7612±0.0040** (0.7583-0.7640)	0.6957±0.0038 ** (0.6927-0.6982)
Random Forest	0.8182±0.0235 (0.8014-0.8350)	0.5110±0.0063** (0.5065-0.5155)	0.5735±0.0083 ** (0.5676-0.5794)	0.9036±0.0049** (0.9001-0.9072)	0.6421±0.0054 ** (0.6378-0.6456)
Decision Tree	0.6017±0.0082 ** (0.5958-0.6076)	0.5589±0.0082** (0.5530-0.5648)	0.5777±0.0075 ** (0.5723-0.5831)	0.7706±0.0058** (0.7665-0.7747)	0.5573±0.0059 ** (0.5531-0.5615)
Light GBM	0.8201±0.0141 * (0.8101-0.8302)	0.5621±0.0108** (0.5543-0.5698)	0.6405±0.0124 ** (0.6316-0.6494)	0.9331±0.0036 (0.9305-0.9356)	0.6756±0.0081 ** (0.6694-0.6810)
DeepFFM	0.3156±0.000** (0.3156-0.3156)	0.3333±0.000** (0.3333-0.3333)	0.3242±0.000** (0.3242-0.3242)	0.5000±0.000** (nan-nan)	0.6667±0.0000 ** (0.6667-0.6667)
RETAIN	0.8138±0.0151 ** (0.8029-0.8246)	0.5930±0.0136 (0.5833-0.6027)	0.6606±0.0167 (0.6487-0.6726)	0.9437±0.0020 ** (0.9422-0.9451)	0.7013±0.0106 * (0.6933-0.7085)
MCNN	0.8447±0.0274 (0.8278-0.8670)	0.5979±0.0272 (0.5775-0.6165)	0.6691±0.0214 (0.6553-0.6859)	0.9364±0.0031 (0.9321-0.9365)	0.7069±0.0078 (0.7041-0.7153)

Note: ** indicates $p < 0.01$; * indicates $p < 0.05$.

Among baselines, Naive Bayes achieved the highest recall of 0.7438 but poor precision of 0.4307, yielding a low F_1 -score of 0.3124. Random Forest showed more balanced performance with a precision of 0.8182 and recall of 0.5110, yielding an F_1 -score of 0.5735. Decision Tree achieved an F_1 -score of 0.5777, while LightGBM showed competitive results with an F_1 -score of 0.6405, precision of 0.8201, and recall of 0.5621. DeepFFM performed poorly with an F_1 -score of 0.3242, precision of 0.3156, and recall of 0.3333, suggesting its factorization machine approach is ill-suited for this clinical prediction task. RETAIN demonstrated the strongest baseline performance with an F_1 -score of 0.6606, precision of 0.8138, and recall of 0.5930, validating the importance of temporal modeling in healthcare prediction.

Figure 2 illustrates AUROC comparisons across all models. RETAIN achieved the highest AUROC of 0.9437, marginally outperforming MCNN's AUROC of 0.9364, followed by LightGBM (0.9331) and Random Forest (0.9036). DeepFFM's AUROC of 0.5000 was equivalent to random classification. Figure 3 presents AUPRC results, where MCNN achieved the highest AUPRC of 0.7069, followed by RETAIN (0.7013). Notably, Naive Bayes achieved a relatively high AUPRC (0.6957) despite its low F_1 -score, underscoring the importance of considering multiple evaluation metrics.

While RETAIN's marginally high AUROC reflects the strength of its reverse-time attention mechanism in capturing sequential dependencies, the MCNN provides practical advantages for large-scale healthcare data analysis through fewer parameters, lower computational overhead, and a multi-channel design that enables simultaneous modeling of heterogeneous clinical modalities.

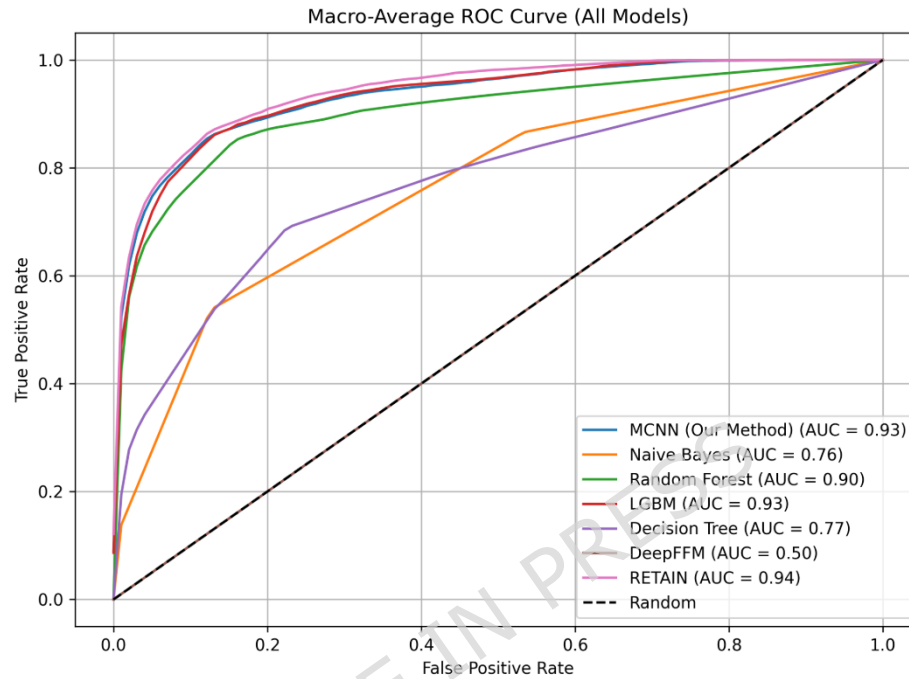


Figure 2. ROC Curves for Lung Cancer Detection Models

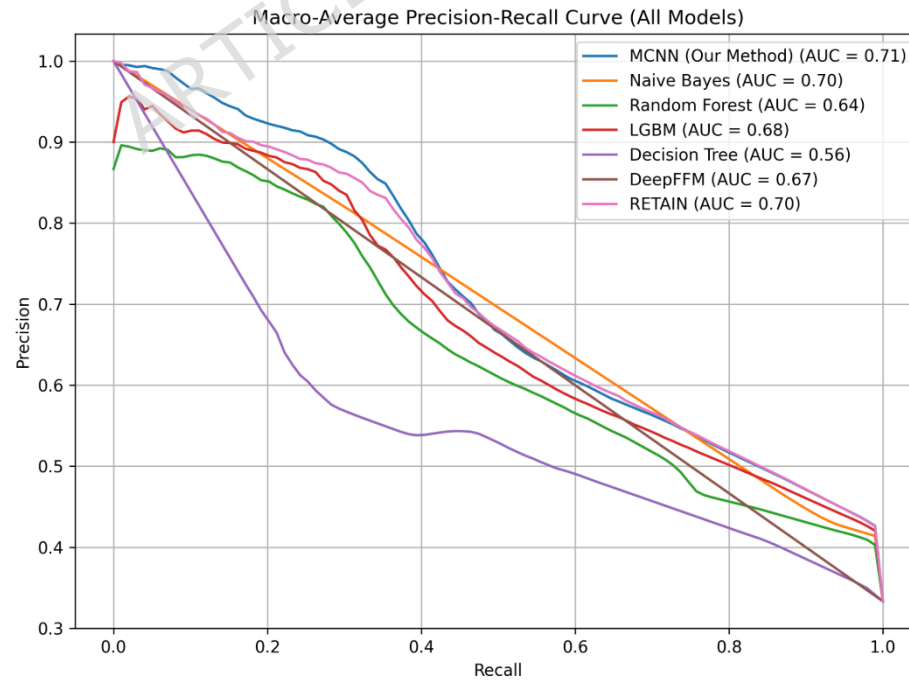


Figure 3. Precision-Recall Curves for Lung Cancer Detection Models

The ablation study confirmed that ICD omission caused the most substantial performance degradation, with the F₁-score dropping to 0.3388, confirming diagnostic codes as the critical modality. Medical order omission showed moderate impact with an F₁-score of 0.5827, while medication omission had the least effect at 0.6314, demonstrating the complementary contribution of each modality. The relative importance hierarchy aligns with clinical intuition: diagnostic history represents the most fundamental component of medical risk assessment, while procedural and pharmaceutical data provide meaningful supplementary signals that collectively enhance predictive performance.

Table 3. Performance impact of individual data modality omission

	Precision	Recall	F ₁ -score	AUROC	AUPRC
MCNN	0.8447±0.0 274 (0.8278- 0.8670)	0.5979±0.0 272 (0.5775- 0.6165)	0.6691±0.0 214 (0.6553- 0.6859)	0.9364±0.0 031 (0.9321- 0.9365)	0.7069±0.0 078 (0.7041- 0.7153)
ICD omission	0.5138±0.0 873 (0.4514- 0.5762)	0.3410±0.0 048 (0.3376- 0.3444)	0.3388±0.0 087 (0.3326- 0.3451)	0.8932±0.0 050 (0.8896- 0.8968)	0.5636±0.0 541 (0.5245- 0.6019)
Medication omission	0.8490±0.0 139 (0.8391- 0.8589)	0.5503±0.0 131 (0.5408- 0.5597)	0.6314±0.0 134 (0.6218- 0.6410)	0.9097±0.0 044 (0.9066- 0.9128)	0.6668±0.0 068 (0.6618- 0.6716)
Medical order omission	0.8363±0.0 232 (0.8197- 0.8528)	0.5038±0.0 134 (0.4942- 0.5134)	0.5827±0.0 141 (0.5725- 0.5928)	0.8950±0.0 028 (0.8930- 0.8970)	0.6075±0.0 075 (0.6022- 0.6128)

To assess temporal generalizability, a calendar-time holdout evaluation was conducted, training on NHIRD data from 1999 to 2012 and evaluating on an independent 2013 cohort without retraining or recalibration, using a separate 100,000-file NHIRD pool. The model maintained stable recall of 0.536 while precision showed a moderate decrease to 0.581, yielding an F₁-score of 0.549. The performance reduction is consistent with the combined effect of a smaller sampling pool, reduced rare-event case counts, and temporal distribution shift between training and validation periods.

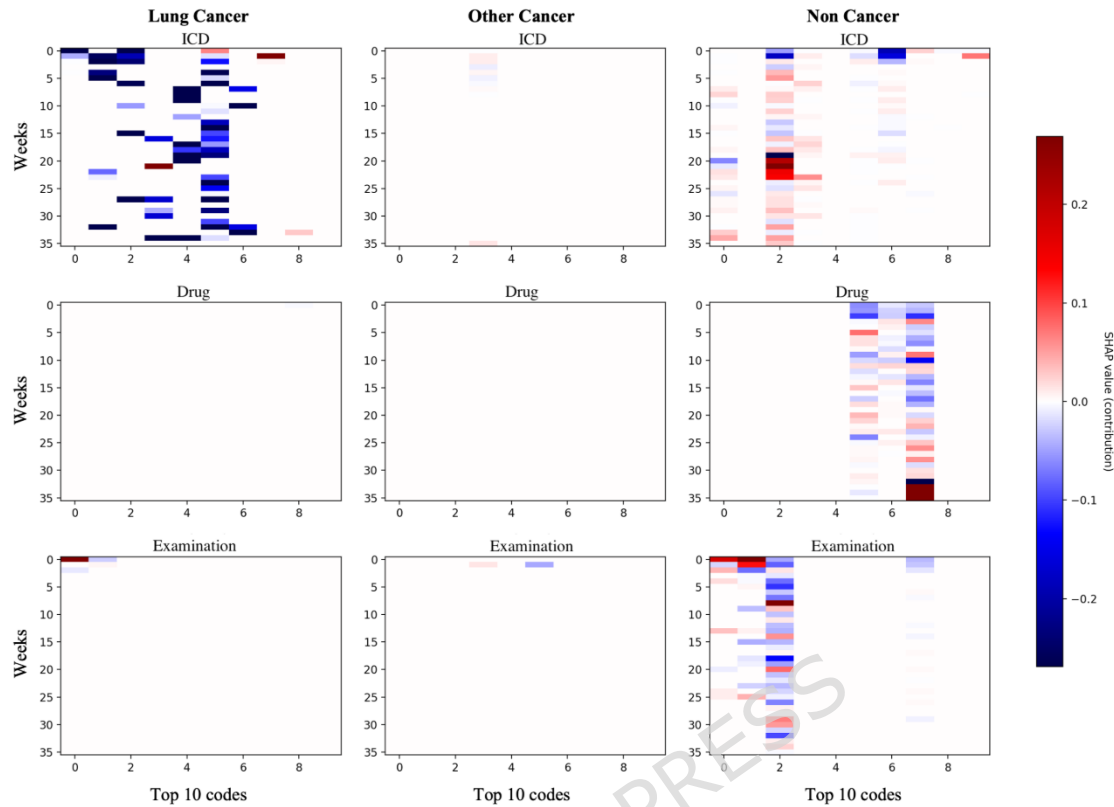


Figure 4. SHAP analysis of temporal feature contributions for lung cancer prediction

Red indicates positive contributions toward lung cancer classification; blue indicates negative contributions.

SHAP analysis was applied to representative cases across all three cohorts, with results for the lung cancer case presented in Figure 4 and corresponding patterns for other cancer and non-cancer cases illustrated in Supplementary Figures 1 and 2. For the lung cancer case, the ICD channel shows predominantly negative contributions (blue coloring) throughout most of the observation period, indicating that the absence of definitive diagnoses during the pre-diagnostic phase provides discriminative information. The examination channel shows strong positive signals (red coloring) early in the observation window, corresponding to cancer-specific procedures preceding formal diagnosis coding. The other cancer case from Supplementary Figure 1 displays more distributed positive contributions across multiple channels, reflecting diverse diagnostic and therapeutic pathways associated with various malignancies. The non-cancer case of Supplementary Figure 2 shows predominant positive contributions from routine healthcare and preventive care codes, validating the model's ability to distinguish normal from pathological engagement patterns. Medication channels across all cases show more concentrated therapeutic interventions in cancer cases compared to distributed routine patterns in non-cancer patients.

The clinical relevance of identified features is presented in Table 4. For the lung cancer cohort, the highest-ranked diagnostic code was malignant

neoplasm of lower lobe bronchus with ICD code 162.5, the most predictive medication was PHYXOL injection, and the top medical order was bronchial biopsy with impression. The other cancer cohort showed distinct patterns with breast cancer diagnoses, cancer-specific medications like CASODEX tablets, and tumor-specific biopsy procedures. The non-cancer cohort was characterized by routine examinations, general medications, and standard diagnostic procedures. The discriminative patterns observed across cohorts are consistent with established clinical pathways, with lung cancer-specific features reflecting diagnostic workup and treatment procedures, while non-cancer features predominantly reflect routine preventive care and general health maintenance. Complete feature rankings are provided in Supplementary Table 1.

Table 4. Top-ranked clinical features for lung cancer detection by patient cohort

Lung cancer cohort						
Rank	ICD code	Diagnostic description	Medication code	Medication description	Medical order code	Medical order description
1	162.5	Malignant neoplasm of lower lobe, bronchus or lung	A0444662E 2	PHYXOL INJECTION 6MG/ML "SINPHAR"	15020B	Bronchial biopsy with impression
2	539	Complications of gastric band procedure	B02088122 1	ETOPOSIDE-TEVA 20MG/ML	27082B	Tissue polypeptide specific antigen (TPS)
3	009.0	Infectious colitis, enteritis and gastroenteritis	A04787222 3	Gemmis Injection 38 mg/ml	29030A	Percutaneous transthoracic lung tissue puncture
Other cancer cohort						
1	174.9	Cancer (Carcinoma) of breast, female; Malignant neoplasm of female breast	B01255410 0	ANDROCUR TABLETS	54042C	ENT (Ear, Nose, Throat) biopsy
2	155.0	Hepatoma	B02280310 0	CASODEX TABLETS 50MG	63010C	Breast tumor tissue biopsy examination
3	154.1	Cancer of rectum	A04369810 0	UFUR CAPSULE	37013B	Alloy module design and fabrication
Non-cancer cohort						
1	V70.0	Routine general medical examination at a health care facility	M13	Billing code for daily dosage (under 6 years old) - legacy code	25004C	Level IV surgical pathology, complex
2	V22.1	Routine postpartum follow-up	A02865650 0	VOREN SUPPOSITORIES 12.5MG (DICLOFENAC)	33072B	CT scan - with/without contrast
3	401.9	Essential hypertension	A04016415 1	ANTI-PHEN SYRUP 24MG/ML "CENTER"	09015C	Serum creatinine

Most predictive clinical features for patient classification by cohort. Shows top 3 features per data modality (diagnostic codes, medications, medical orders) ranked by discriminative strength. Complete feature rankings with discriminative weights are provided in Supplementary Table 1.

Discussion

The MCNN achieved a superior F₁-score of 66.91% with exceptional precision of 84.47%, demonstrating the potential of longitudinal clinical data analysis for oncological risk assessment. The poor performance of DeepFFM, with an F₁-score of only 32.42%, suggests that general-purpose deep learning architectures are ill-suited for clinical prediction tasks requiring temporal and multi-modal integration [28], while RETAIN's competitive performance (F₁-score of 66.06%) validates the importance of temporal modeling in healthcare prediction [29]. These results are particularly noteworthy in the context of lung cancer screening, where the balance between sensitivity and specificity directly impacts patient outcomes and resource use. While RETAIN's marginally higher AUROC (0.9437 vs. 0.9364) reflects the strength of its reverse-time attention mechanism, the MCNN's fewer parameters, lower computational overhead, and multi-channel design offer practical advantages for large-scale administrative database deployment.

The ablation findings confirm the multi-modal framework's value with diagnostic codes omission causing the most severe performance drop, with the F₁-score falling to 33.88%, followed by medical order omission at 58.27% and medication omission at 63.14%. Building on these findings, the SHAP analysis revealed that the predominantly negative ICD channels contributions for lung cancer cases reflect the model's ability to identify pre-diagnostic phases where patients undergo cancer-related procedures before receiving definitive diagnoses in administrative coding. The examination channel patterns, with strong positive signals corresponding to cancer-specific procedures preceding formal diagnosis coding, demonstrate the model's capacity to identify patients during critical pre-diagnostic evaluation phases when intervention potential is highest.

An interesting observation from the SHAP analysis was that earlier periods in the observation window sometimes exhibited limited disease-specific diagnostic activity among patients who were later diagnosed with lung cancer. This phenomenon can be interpreted as a form of "clinical silence" during the pre-diagnostic phase. Lung cancer often develops gradually, with subtle or non-specific symptoms that may not immediately lead to a definitive diagnosis, while explicit diagnostic signals remain absent. Within a longitudinal modeling framework, the absence of strong diagnostic signals in earlier months becomes informative when contrasted with the emergence of respiratory-related examinations or procedures closer to the diagnosis date, capturing the temporal evolution of healthcare utilization patterns preceding diagnosis. The SHAP patterns reflect the model's learned decision boundaries applied to representative individual cases, where each patient's full longitudinal profile serves as input, and were selected to illustrate trajectories consistent with those commonly observed within each cohort. Some predictors may also reflect indirect healthcare utilization patterns or diagnostic processes prior to cancer detection rather than direct causal risk factors, as recorded diagnostic codes in administrative claims data

may function as proxies for broader clinical workups leading to a final diagnosis.

The identification of ICD code 162.5 as the most discriminative diagnostic feature aligns with established epidemiological patterns [32], the prominence of etoposide-based agents among top pharmaceutical features directly reflects the lung cancer therapeutic landscape [33], and the emergence of bronchial biopsy procedures as the most discriminative medical orders aligns with established clinical diagnostic pathways [34]. Regarding the computational efficiency, systematic feature reduction estimated storage requirements at approximately 1,184 GB for the full ICD feature space ($523,539 \text{ patients} \times 36 \text{ months} \times 16,870 \text{ ICD codes} \times 4 \text{ bytes} \times 1 \text{ table/patient}$), reduced to approximately 2.11 GB for the feature-reduced approach ($523,539 \text{ patients} \times 36 \text{ months} \times 10 \text{ ICD codes} \times 4 \text{ bytes} \times 3 \text{ tables/patient}$), representing a 99.8% reduction that illustrates the scalability advantage of the proposed dimensionality reduction strategy [35].

Current LDCT screening faces substantial barriers including high costs, limited accessibility, and false positive rates up to 24% [6]. The proposed MCNN offers a complementary strategy to enhance screening efficiency by identifying high-risk individuals for imaging studies [36]. In a practical deployment scenario, routinely collected structured healthcare data could be extracted automatically from the EHR systems over a predefined retrospective period, and the model could generate patient-level risk scores presented to clinicians as a non-interruptive decision-support signal, such as dashboard flags in primary care or outpatient settings, intended to support rather than replace physician judgment. Operating thresholds could be calibrated according to population prevalence and institutional screening capacity, enabling the model to function as a population-level risk stratification tool. Nevertheless, prospective implementation studies are needed to determine appropriate alert thresholds and evaluate real-world impact. Future evaluations should also assess model calibration and clinical utility using calibration curves, Brier scores, and decision curve analysis.

Beyond methodological considerations, it is important to situate the proposed model relative to established lung cancer risk tools. Models such as PLCOm2012 [37] and LCRAT [38] incorporate individual-level smoking history, pack-years, and other behavioral variables that are not available in administrative claims databases, making direct numerical comparison with the proposed MCNN methodologically inappropriate given their fundamentally different input spaces. The MCNN is therefore not intended as a replacement for these risk factor-based tools but rather as a complementary claims-based screening instrument, leveraging routinely collected structured healthcare data to identify high-risk individuals from the broader healthcare-utilizing population for further evaluation. Validation in independent cohorts and integration with risk factor data remain important directions for future work.

Limitations

However, several limitations of this study should be acknowledged. First, reliance on NHIRD data from 2013 may not fully reflect more recent advances in lung cancer diagnosis and treatment [39]. Second, administrative claims data do not capture important lifestyle and environmental risk factors, including smoking history and occupational exposures, which limits the model's ability to incorporate established risk predictors [40]. Third, lung cancer cases were identified using administrative diagnostic codes. Because registry or pathology linkage was not available in the current dataset, some degree of misclassification cannot be completely excluded. Fourth, the geographic specificity of our study population raises questions about generalizability to populations with different demographic characteristics and healthcare systems [41]. While a calendar-time holdout evaluation using an independent 2013 NHIRD pool demonstrated temporal generalizability within the same healthcare system, this does not constitute validation in an independent external cohort, and future studies in geographically diverse populations remain essential. Additionally, detailed characteristics of patients excluded due to insufficient healthcare encounters were not available for comparison, which limits our ability to fully characterize the representativeness of the study sample. Future validation studies using more recent and internationally diverse datasets will be important for assessing broader applicability. Moreover, the model is designed to identify high-risk individuals within healthcare-utilizing populations rather than asymptomatic individuals without healthcare interactions.

CONCLUSION

This study demonstrates the potential of integrating longitudinal clinical data with artificial intelligence for early lung cancer risk prediction. The proposed MCNN achieved strong predictive performance, with an F₁-score of 66.91% and precision of 84.47%, highlighting the feasibility of AI-driven oncological risk assessment using administrative healthcare data. By leveraging three years of diagnostic codes, medication records, and medical order information, the multi-modal framework was able to identify high-risk patients during the critical pre-diagnostic phase when early intervention remains possible.

Ablation analyses further demonstrated the complementary contributions of multiple data modalities, with diagnostic information serving as the primary predictive signal while medication and medical order data provided additional context that improved risk assessment. The SHAP-based interpretability analysis showed that the model captures temporal healthcare utilization patterns consistent with known clinical pathways preceding lung cancer diagnosis, supporting the clinical plausibility of the proposed framework. In addition, systematic feature selection substantially improved

computational efficiency, reducing memory requirements by 99.8% while maintaining predictive performance.

The present findings should be interpreted as exploratory, as the model was developed and validated within a single national healthcare database without geographic external validation and without access to key risk factors such as smoking history. Overall, the MCNN framework provides a practical approach for leveraging routinely collected healthcare data to support risk stratification in lung cancer screening programs. Rather than replacing existing screening strategies, such models may serve as complementary tools to help identify individuals who would benefit most from further diagnostic evaluation. Future research should focus on external validation using diverse healthcare systems, integration of additional risk factors such as lifestyle and environmental exposures, and the development of explainable AI techniques specifically designed for clinical applications [42]. These efforts will be essential for translating AI-based risk prediction models into scalable and clinically useful decision-support tools.

DATA AVAILABILITY

The datasets generated and/or analyzed during this study are available from the corresponding author upon reasonable request. The model code and implementation details are also available from the corresponding author upon reasonable request, as the framework is part of ongoing research and development.

FUNDING

This research was supported by the National Science and Technology Council of Taiwan, grant number NSTC 113-2221-E-038-019-MY3 and NSTC 114-2410-H-038-034-MY3.

COMPETING INTERESTS

The authors have no conflict of interest to declare.

CONTRIBUTORS

Chia-Hui Chien: Writing – original draft, Validation, Investigation, Formal analysis, Data curation, Methodology. Shih-Chuan Chang: Validation, Formal analysis, Methodology. Yung-Chun Chang: Writing – original draft, review & editing, Resources, Validation, Investigation, Formal analysis, Conceptualization, Supervision, Project administration. Yu-Chuan Li: Writing – original draft, review & editing, Resources, Conceptualization, Methodology, Project administration.

References

1. Bray, F., et al., *Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA: a cancer journal for clinicians, 2024. **74**(3): p. 229-263.
2. Welfare., T.M.o.H.a. *Cancer Registry Annual Report*. 2021; Available from: <https://www.mohw.gov.tw/>.

3. Nam, B., et al., *Lung Cancer Screening: Early Detection Decreases Mortality*. Delaware Journal of Public Health, 2024. **10**(3): p. 22.
4. Jonas, D.E., 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**(10): p. 971-987.
5. Team, N.L.S.T.R., *Reduced lung-cancer mortality with low-dose computed tomographic screening*. New England Journal of Medicine, 2011. **365**(5): p. 395-409.
6. Oudkerk, M., et al., *Lung cancer LDCT screening and mortality reduction—evidence, pitfalls and future perspectives*. Nature reviews Clinical oncology, 2021. **18**(3): p. 135-151.
7. Esteva, A., et al., *Dermatologist-level classification of skin cancer with deep neural networks*. nature, 2017. **542**(7639): p. 115-118.
8. Topol, E.J., *High-performance medicine: the convergence of human and artificial intelligence*. Nature medicine, 2019. **25**(1): p. 44-56.
9. Chen, A., et al., *Development of lung cancer risk prediction machine learning models for equitable learning health system: retrospective study*. JMIR AI, 2024. **3**: p. e56590.
10. Yaqoob, A., R. Musheer Aziz, and N.K. Verma, *Applications and techniques of machine learning in cancer classification: a systematic review*. Human-centric intelligent systems, 2023. **3**(4): p. 588-615.
11. Litjens, G., et al., *A survey on deep learning in medical image analysis*. Medical image analysis, 2017. **42**: p. 60-88.
12. Rajpurkar, P., et al., *Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning*. arXiv preprint arXiv:1711.05225, 2017.
13. Xiao, C., E. Choi, and J. Sun, *Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review*. Journal of the American Medical Informatics Association, 2018. **25**(10): p. 1419-1428.
14. Swinckels, L., et al., *The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: scoping review*. Journal of medical Internet research, 2024. **26**: p. e48320.
15. Miotto, R., et al., *Deep learning for healthcare: review, opportunities and challenges*. Briefings in bioinformatics, 2018. **19**(6): p. 1236-1246.
16. Shickel, B., et al., *Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis*. IEEE journal of biomedical and health informatics, 2017. **22**(5): p. 1589-1604.
17. Levi, M., et al., *Machine learning computational model to predict lung cancer using electronic medical records*. Cancer Epidemiology, 2024. **92**: p. 102631.
18. Lin, L.-Y., et al., *Data resource profile: the national health insurance research database (NHIRD)*. Epidemiology and health, 2018. **40**: p. e2018062.
19. Hsieh, C.-Y., et al., *Taiwan's national health insurance research database: past and future*. Clinical epidemiology, 2019: p. 349-358.
20. Kao, W.H., et al., *Validity of cancer diagnosis in the National Health Insurance database compared with the linked National Cancer Registry in Taiwan*. Pharmacoepidemiology and drug safety, 2018. **27**(10): p. 1060-1066.
21. Organization, W.H., *WHO Collaborating Centre for Drug Statistics Methodology, ATC/DDD Index 2023*. 2023. 2023.

22. Yang, Y. and J.O. Pedersen. *A comparative study on feature selection in text categorization*. in *icml*. 1997. Citeseer.
23. Yaqoob, A., N.K. Verma, and R.M. Aziz, *Improving breast cancer classification with mRMR+ SS0+ WSVM: a hybrid approach*. Multimedia Tools and Applications, 2025. **84**(22): p. 26041-26066.
24. Javidi, H., et al., *Identification of robust deep neural network models of longitudinal clinical measurements*. NPJ Digital Medicine, 2022. **5**(1): p. 106.
25. Wang, H.-H., et al., *Assessment of deep learning using nonimaging information and sequential medical records to develop a prediction model for nonmelanoma skin cancer*. JAMA dermatology, 2019. **155**(11): p. 1277-1283.
26. Yeh, M.C.-H., et al., *Artificial intelligence-based prediction of lung cancer risk using nonimaging electronic medical records: deep learning approach*. Journal of medical Internet research, 2021. **23**(8): p. e26256.
27. Catling, F.J. and A.H. Wolff, *Temporal convolutional networks allow early prediction of events in critical care*. Journal of the American Medical Informatics Association, 2020. **27**(3): p. 355-365.
28. Guo, H.T., Ruiming; Ye, Yunming; Li, Zhenguo; He, Xiugiang. *DeepFM: a factorization-machine based neural network for CTR prediction*. in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*. 2017.
29. Choi, E., et al., *Retain: An interpretable predictive model for healthcare using reverse time attention mechanism*. Advances in neural information processing systems, 2016. **29**.
30. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions*. Advances in neural information processing systems, 2017. **30**.
31. Alkhanbouli, R., et al., *The role of explainable artificial intelligence in disease prediction: a systematic literature review and future research directions*. BMC medical informatics and decision making, 2025. **25**(1): p. 110.
32. Travis, W.D., et al., *The 2015 World Health Organization classification of lung tumors: impact of genetic, clinical and radiologic advances since the 2004 classification*. Journal of thoracic oncology, 2015. **10**(9): p. 1243-1260.
33. Hanna, N., et al., *Systemic therapy for stage IV non-small-cell lung cancer: American Society of Clinical Oncology clinical practice guideline update*. Journal of Clinical Oncology, 2017. **35**(30): p. 3484-3515.
34. Rivera, M.P., A.C. Mehta, and M.M. Wahidi, *Establishing the diagnosis of lung cancer: Diagnosis and management of lung cancer: American College of Chest Physicians evidence-based clinical practice guidelines*. Chest, 2013. **143**(5): p. e142S-e165S.
35. Chen, J.H. and S.M. Asch, *Machine learning and prediction in medicine—beyond the peak of inflated expectations*. The New England journal of medicine, 2017. **376**(26): p. 2507.
36. Chandran, U., et al., *Machine learning and real-world data to predict lung cancer risk in routine care*. Cancer Epidemiology, Biomarkers & Prevention, 2023. **32**(3): p. 337-343.
37. Tammemägi, M.C., et al., *Selection criteria for lung-cancer screening*. New England Journal of Medicine, 2013. **368**(8): p. 728-736.

38. Katki, H.A., et al., *Development and validation of risk models to select ever-smokers for CT lung cancer screening*. *Jama*, 2016. **315**(21): p. 2300-2311.
39. Hirsch, F.R., et al., *Lung cancer: current therapies and new targeted treatments*. *The Lancet*, 2017. **389**(10066): p. 299-311.
40. Malhotra, J., et al., *Risk factors for lung cancer worldwide*. *European Respiratory Journal*, 2016. **48**(3): p. 889-902.
41. Torre, L.A., et al., *Global cancer statistics, 2012*. *CA: a cancer journal for clinicians*, 2015. **65**(2): p. 87-108.
42. Rajkomar, A., J. Dean, and I. Kohane, *Machine learning in medicine*. *New England Journal of Medicine*, 2019. **380**(14): p. 1347-1358.

ARTICLE IN PRESS