

Early detection of female-specific cancers using longitudinal healthcare records with a multichannel convolutional neural network

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

To cite: Chien C-H, Chang S-C, Chang Y-C, *et al.* Early detection of female-specific cancers using longitudinal healthcare records with a multichannel convolutional neural network. *BMJ Health Care Inform* 2026;**33**:e101874. doi:10.1136/bmjhci-2025-101874

► Additional supplemental material is published online only. To view, please visit the journal online (<https://doi.org/10.1136/bmjhci-2025-101874>).

Received 14 October 2025
Accepted 27 December 2025



© 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.

¹Computer Science, Middlesex University, London, UK

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

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

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

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

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

ABSTRACT

Objectives Female-specific cancers, including breast, ovarian, cervical and uterine malignancies, lack comprehensive early detection approaches, particularly for ovarian and endometrial cancers where effective population-level screening remains limited. This study aimed to develop and validate a computational method for early detection of female-specific cancers using longitudinal healthcare records.

Methods We developed a multichannel convolutional neural network (MCNN) to analyse 36-month pre-diagnostic healthcare records from Taiwan's National Health Insurance Research Database. The study included 19 954 female patients (596 cancer cases, 19 358 controls) from 1999 to 2013. Log-likelihood ratio feature selection identified top 10 features across three data modalities (diagnostic codes, medications, medical orders). The six-channel architecture processed temporal patterns through stratified 10-fold cross-validation, with performance compared against nine baseline algorithms.

Results MCNN achieved superior balanced performance with Macro-F₁ score of 0.8443, precision of 0.9135 and recall of 0.7978, outperforming traditional machine learning and deep learning approaches. Feature analysis revealed clinically relevant patterns including tamoxifen therapy, immunohistochemical procedures and cancer-specific diagnostic codes. SHapley Additive exPlanations (SHAP) interpretability analysis demonstrated the model's ability to identify pre-diagnostic phases through temporal healthcare utilisation patterns. Systematic feature selection reduced computational requirements by over 99%, enabling validation on Taiwan's population-scale National Health Insurance Research Database (NHIRD).

Discussion The multichannel deep learning approach enables unified early detection across four female cancer types using routine administrative data, addressing detection gaps for ovarian and endometrial cancers while providing complementary risk stratification for existing screening programmes.

Conclusion Clinical implementation through electronic health record (EHR) integration offers practical pathways for accessible cancer risk assessment during routine healthcare encounters.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Established screening programmes exist for breast and cervical cancers, but systematic early detection for ovarian and endometrial cancers remains ineffective due to asymptomatic early stages and lack of reliable population-level screening tools.

WHAT THIS STUDY ADDS

⇒ Multichannel deep learning simultaneously detects four female-specific cancer types during pre-diagnostic phases using routine administrative data, with 99% reduced computational requirements enabling population-scale validation and clinical deployment.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ This approach provides the first systematic computational method for ovarian and endometrial cancer screening and enables electronic health record (EHR) integration for seamless risk assessment during routine healthcare encounters.

INTRODUCTION

Female-specific cancers, including breast, ovarian, cervical and uterine malignancies, represent a substantial global health burden with significant impact on women's mortality and healthcare systems. According to Taiwan's Ministry of Health and Welfare statistics, these four cancer types consistently rank among the leading causes of cancer-related deaths in women, making them particularly suitable for population-based early detection research.¹ Early-stage identification dramatically improves survival outcomes across all these cancer types, yet current detection approaches demonstrate significant limitations. Established screening programmes show variable effectiveness across different cancer types. Cervical cancer screening through Pap smears has reduced incidence rates by 60%–80% in populations

with robust programmes, while mammography screening demonstrates 20%–25% mortality reduction for breast cancer.^{2,3} However, systematic early detection for ovarian and endometrial cancers remains challenging due to asymptomatic early stages and lack of effective population-level screening tools.⁴ Ovarian cancer, often presenting with non-specific symptoms, is detected at early stages in only 15% of cases.⁵ Current biomarker approaches using CA-125 and transvaginal ultrasound show limited effectiveness due to high false-positive rates and insufficient sensitivity.⁶ These limitations underscore the need for novel approaches that can identify cancer risk patterns before traditional screening methods become effective.

The evolution of comprehensive healthcare databases has created new opportunities for early detection research. Taiwan's National Health Insurance Research Database (NHIRD) represents one of the most comprehensive longitudinal healthcare datasets globally, covering over 99% of Taiwan's population with detailed medical records spanning multiple decades.⁷ This database captures healthcare utilisation patterns, prescription histories and procedural records that may contain subtle signals preceding cancer diagnosis. Recent investigations suggest that healthcare utilisation patterns may change months or years before formal cancer diagnosis, indicating detectable preclinical phases.⁸ Advances in computational methods, particularly machine learning and artificial intelligence, have opened new avenues for analysing complex healthcare datasets to identify multi-dimensional patterns impossible to detect through traditional statistical methods.^{9,10} The economic implications are substantial as late-stage cancer treatment costs often exceed early-stage treatment by factors of three to five.¹¹

This study investigates computational approaches for early detection of female-specific cancers using longitudinal healthcare data from Taiwan's NHIRD. Our approach analyses 36 months of pre-diagnostic healthcare utilisation patterns to identify subtle changes preceding clinical recognition of malignancy. The use of first hospitalisation as the diagnostic timepoint provides a clinically meaningful anchor capturing the transition from outpatient management to intensive medical intervention. The primary objectives are: first, to develop and validate computational methods for processing longitudinal healthcare data for early cancer detection; second, to evaluate feasibility and performance characteristics for clinical implementation; and third, to assess potential utility as decision support tools in healthcare environments. We hypothesise that advanced computational methods applied to longitudinal healthcare data will demonstrate competitive performance while offering practical advantages for clinical deployment and improved patient outcomes.

This study provides three key contributions to early cancer detection research. First, we present a unified deep learning model that detects breast, ovarian, cervical and uterine cancers concurrently using population-scale administrative health data. Second, the proposed

multichannel convolutional neural network (MCNN) integrates three heterogeneous modalities (diagnoses, medications and medical orders) and captures temporal interactions that prior models have not leveraged. Third, we develop a highly efficient feature-selection strategy that reduces computational burden by over 99%, enabling both population-level validation and practical deployment in clinical environments. Together, these contributions advance the methodological and translational foundations for scalable early detection of female-specific cancers.

MATERIALS AND METHODS

Study population and data preprocessing

This study used data from NHIRD, covering over 99% of Taiwan's population.¹² All data are deidentified and maintained by the Health and Welfare Data Science Center.⁷ This study was approved by the Institutional Review Board of Taipei Medical University, TMU-JIRB No.: N202405151, with informed consent waived due to fully anonymised data. The study population was derived from NHIRD files covering 1 January 1999 to 31 December 2013, restricted to female patients. The cancer cohort comprised patients with confirmed diagnoses of breast cancer (ICD-9: 174.xx), ovarian cancer (ICD-9: 183.xx), cervical cancer (ICD-9: 180.xx) or uterine cancer (ICD-9: 179.xx, 182.xx), requiring at least one inpatient hospitalisation record. The non-cancer cohort consisted of female patients without any cancer diagnoses during the observation period.

Inclusion criteria required continuous insurance enrolment for at least 36 months prior to the index date, age ≥ 18 years and at least one healthcare encounter annually. The 36-month observation period was calculated retrospectively from the first cancer-related hospitalisation date for patients with cancer, and from the last hospitalisation date for non-cancer patients. This observation window was selected based on prior literature indicating that healthcare utilisation changes may manifest 2–3 years before cancer diagnosis.¹³ The final study population consisted of 19954 female patients (596 cancer cases, 19358 controls) as shown in [table 1](#).

Feature selection used log-likelihood ratio (LLR) methodology to identify clinically relevant predictors from diagnostic codes (International Classification of Diseases (ICD)-9), medication prescriptions (Anatomical Therapeutic Chemical (ATC) classification) and medical order codes.^{14,15} The top 10 features with highest LLR scores were selected from each modality, resulting in 30 primary features. Temporal encoding mapped each feature across 36 monthly intervals. LLR feature selection was performed on the entire dataset prior to cross-validation to ensure consistent feature sets across evaluation folds; this methodological choice is discussed as a limitation. Details of the LLR formulation are provided in online supplemental methods.

Table 1 Demographic and clinical characteristics of study population

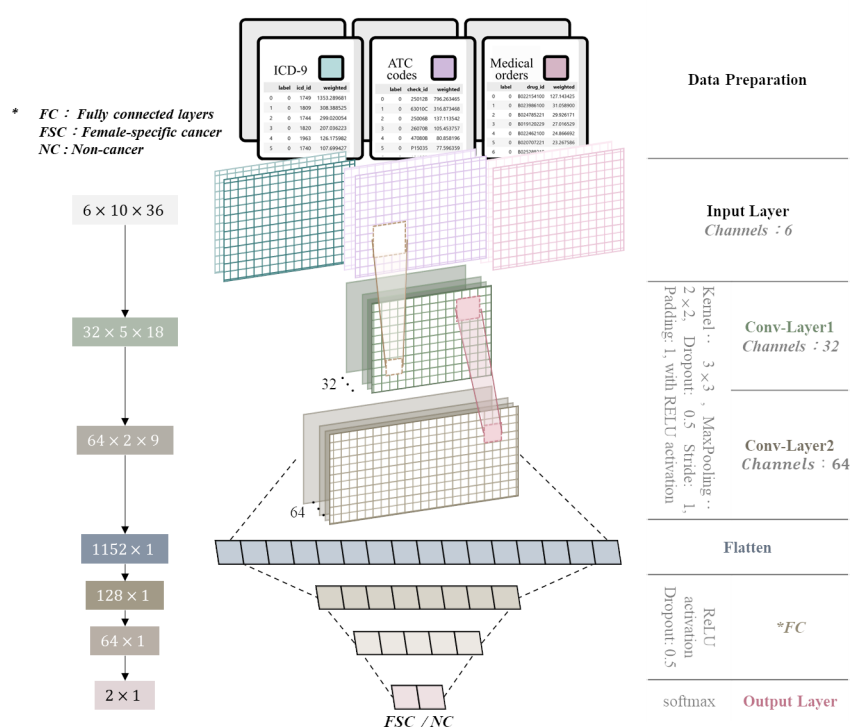
Variables	Cohort population (n=19954)		P value
	Female cancer	Non-cancer	
Patient number	596	19 358	
Breast cancer	418 (70.1%)	–	
Cervical cancer	90 (15.1%)	–	
Ovarian cancer	40 (6.7%)	–	
Uterine cancer	61 (10.2%)	–	
Total cancer diagnosis	617	–	
Age, mean (SD), year	53.6 (12.03)	45.3 (18.21)	<0.001
18–34	20 (3.4%)	7325 (37.8%)	
34–64	483 (81.0%)	8537 (44.1%)	
>65	93 (15.6%)	3496 (18.1%)	
Monthly insured salary, mean (SD), NTD	17 253.1 (19 394.53)	18 918.1 (18 945.43)	<0.01
<19 999	304 (51.0%)	8832 (45.6%)	
20 000–39 999	217 (36.4%)	8247 (42.6%)	
>40 000	75 (12.6%)	2279 (11.8%)	

Some patients had multiple cancer types; percentages may exceed 100%.
NTD stands for New Taiwan Dollar, the currency used in Taiwan.

CNN architecture and evaluation

The proposed MCNN processes diagnostic, medication and medical order data streams simultaneously through a unified framework (figure 1). The input is modelled as a four-dimensional tensor with 6 channels (three data modalities for both cohorts), 10 features per channel and 36 monthly intervals. The architecture consists of two convolutional blocks with 32 and 64 filters, respectively,

each using 3×3 kernels followed by ReLU activation and 2×2 max-pooling. The output is flattened and passed through a fully connected layer with sigmoid activation for binary classification. This hierarchical design captures both local temporal variations and broader disease progression patterns. Detailed architectural specifications and mathematical formulations are provided in online supplemental methods.

**Figure 1** Proposed multichannel convolutional neural network architecture.

Model evaluation employed stratified 10-fold cross-validation with hyperparameter optimisation targeting F1-score maximisation. Performance metrics included precision, recall, F1-score, area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). Comparative analysis included probabilistic methods (Multinomial Naive Bayes), tree-based approaches (Decision Tree, Random Forest), distance-based methods (K-Nearest Neighbours), linear models (Logistic Regression), gradient boosting (XGBoost, LightGBM) and deep learning architectures (CNN, ConvBERT).^{16–19} SHAP analysis quantified individual feature contributions to predictions.²⁰

RESULTS

Our MCNN achieved superior balanced performance for early detection of breast, ovarian, cervical and uterine cancers using 36-month longitudinal healthcare records. **Table 2** compares 10 algorithms across multiple evaluation metrics using stratified 10-fold cross-validation on 19954 female patients (596 cancer cases, 19358 controls). The MCNN achieved the highest Macro-F1 score of 0.8443 with precision of 0.9135 and recall of 0.7978. The model achieved the highest F1 scores for both cancer cases (0.6966) and non-cancer cases (0.9920). The AUROC of 0.9419 and AUPRC of 0.7379 confirm strong discriminative ability. Among baseline methods, gradient boosting approaches (XGBoost: 0.8340, LightGBM: 0.8201) and ConvBERT (0.8251) showed competitive performance, while Random Forest achieved highest precision (0.9715) but poor recall (0.5763).

Table 3 presents top-ranked clinical features revealing clinically intuitive patterns. For patients with cancer, highest-ranked features included malignant neoplasm diagnostic codes (ICD 174.9, 180.9, 174.4), cancer therapeutics (tamoxifen, antiemetics) and diagnostic procedures (immunohistochemical stains, mastectomy, frozen sections). The non-cancer cohort exhibited routine healthcare patterns including benign breast conditions and pregnancy supervision codes. Complete feature rankings are provided in online supplemental table 1.

SHAP analysis examined temporal feature contributions across the 36-month observation period (**figure 2**). Diagnostic code channels showed predominantly negative or neutral contributions during months 12–24, reflecting absence of cancer-related activity, with positive contributions emerging in later months. For medication features, positive SHAP values concentrated in final months corresponding to cancer therapeutics initiation. For examination features, pronounced peaks corresponded to confirmatory diagnostic procedures. These patterns indicate that the model identifies patients during active diagnostic evaluation and early treatment phases rather than truly pre-symptomatic periods, demonstrating utility in systematically identifying patients undergoing cancer evaluation from administrative data patterns.

DISCUSSION

The MCNN achieved balanced performance across breast, ovarian, cervical and uterine cancer detection with Macro-F1 score of 0.8443 and precision of 0.9135, demonstrating the feasibility of unified predictive modelling for multiple female malignancies.²¹ This multicancer approach differs from single-disease prediction models and addresses the clinical need for comprehensive women's cancer risk assessment using routine healthcare data. Identifying these cancers during pre-diagnostic phases addresses critical gaps in women's cancer surveillance. While established screening programmes exist for breast and cervical cancers, our approach provides complementary risk stratification that could identify high-risk individuals between routine screening intervals. For ovarian and endometrial cancers, where effective population-level screening remains elusive, our model offers the first systematic approach to early detection using administrative healthcare data.⁴ The feature analysis revealed clinically meaningful patterns that align with established cancer care pathways. The identification of tamoxifen (Nolvadex) and chemotherapy-related medications among top-ranked pharmaceutical features reflects the therapeutic landscape of female cancer treatment.²² The prominence of immunohistochemical stains and surgical pathology procedures among discriminative medical orders corresponds to standard diagnostic protocols for cancer confirmation.²³ These findings validate the model's ability to recognise clinically relevant healthcare utilisation patterns rather than spurious correlations, supporting its potential clinical utility.

The SHAP interpretability analysis reveals unique temporal patterns in female cancer progression, particularly the identification of patients in pre-diagnostic evaluation phases through absence patterns in diagnostic coding combined with increased procedural activity.²⁴ This finding provides mechanistic insight into healthcare utilisation changes that precede formal diagnosis, differing from traditional biomarker-based early detection approaches. The computational efficiency achieved through systematic feature selection represents a critical methodological contribution enabling population-scale validation and clinical translation. Reducing memory requirements by over 99% compared with full ICD code processing (from 1163.44 GB to 2.07 GB) while maintaining superior predictive performance addresses a fundamental barrier in healthcare artificial intelligence (AI) deployment.²⁵ This efficiency enables: (1) validation on Taiwan's comprehensive NHIRD covering 99% of the population, impossible with full-feature approaches; (2) deployment on standard clinical workstations without specialised hardware; and (3) real-time risk assessment during patient encounters. The lightweight architecture bridges the gap between research demonstrations and practical clinical tools, representing a crucial step towards accessible, scalable early detection systems that can be implemented across diverse healthcare settings.

Table 2 Performance comparison of female-specific cancer detection models

	Precision	Recall	Macro-F ₁	F ₁ _label positive	F ₁ _label negative	AUROC	AUPRC	Specificity	PPV	NPV
Multinomial NB	0.7214	0.8761	0.7759	0.5703 (0.5446–0.5960)	0.9815 (0.9798–0.9833)	0.9312 (0.9181–0.9443)	0.5999 (0.5720–0.6278)	0.9702	0.4498	0.9931
Decision Tree	0.8433	0.7980	0.8185	0.6472 (0.6186–0.6758)	0.9899 (0.9891–0.9907)	0.8344 (0.8130–0.8558)	0.5213 (0.4814–0.5612)	0.9919	0.6988	0.9879
RF	0.9715	0.5763	0.6249	0.2627 (0.2263–0.2991)	0.9870 (0.9866–0.9875)	0.9470 (0.9355–0.9584)	0.7112 (0.6923–0.7402)	0.9998	0.9684	0.9746
KNN	0.8787	0.7519	0.8008	0.6115 (0.5746–0.6483)	0.9901 (0.9892–0.9910)	0.8410 (0.8115–0.8705)	0.5806 (0.5318–0.6294)	0.9953	0.7727	0.9851
LR	0.9139	0.7029	0.7681	0.5465 (0.5022–0.5908)	0.9898 (0.9890–0.9905)	0.9361 (0.9246–0.9476)	0.6886 (0.6579–0.7193)	0.9976	0.8458	0.9821
XGB	0.8865	0.7955	0.8340	0.6767 (0.6444–0.7090)	0.9913 (0.9905–0.9921)	0.9248 (0.9057–0.9439)	0.6896 (0.6490–0.7302)	0.9950	0.7854	0.9876
LGBM	0.8882	0.7747	0.8201	0.6493 (0.6123–0.6863)	0.9909 (0.9901–0.9917)	0.9304 (0.9129–0.9478)	0.6780 (0.6355–0.7205)	0.9955	0.7901	0.9864
CNN	0.8829	0.7453	0.7953	0.6005 (0.5673–0.6338)	0.9900 (0.9892–0.9908)	0.9136 (0.8982–0.9290)	0.6446 (0.6123–0.6769)	0.9954	0.7812	0.9846
ConvBERT	0.9191	0.7737	0.8251	0.6587 (0.6023–0.7152)	0.9914 (0.9903–0.9926)	0.9461 (0.9360–0.9562)	0.7263 (0.6843–0.7682)	0.9966	0.8518	0.9863
MCNN	0.9135	0.7978	0.8443	0.6966 (0.6727–0.7204)	0.9920 (0.9915–0.9926)	0.9419 (0.9287–0.9551)	0.7379 (0.7082–0.7677)	0.9963	0.8392	0.9878

Performance metrics evaluated using stratified 10-fold cross-validation on 19954 female patients (596 cancer cases, 19358 controls).

Bold values indicate best performance for each metric.

F₁_label positive=F₁-score for cancer class; F₁_label negative=F₁-score for non-cancer class.

CNN, convolutional neural network; ConvBERT, Convolutional BERT; KNN, K-Nearest Neighbours; LGBM, LightGBM; LR, Logistic Regression; MCNN, multichannel convolutional neural network; Multinomial NB, Multinomial Naive Bayes; NPV, negative predictive value; PPV, positive predictive value; RF, Random Forest; XGB, XGBoost.

Table 3 Top-ranked clinical features for female-specific cancer detection

Female-specific cancer cohort						
Rank	ICD code	Diagnostic description	Medication code	Medication description	Medical order code	Medical order description
1	174.9	Malignant neoplasm of breast (female), unspecified	B022154100	Nolvadex tablets 10 mg	25012B	Immunohistochemical stains, each antibody
2	180.9	Malignant neoplasm of cervix uteri, unspecified site	B023986100	Emend capsules 125 mg	63010C	Modified radical mastectomy—unilateral
3	174.4	Malignant neoplasm of upper-outer quadrant of female breast	B024785221	Aloxi solution for injection	25006B	Frozen section
Non-cancer cohort						
1	611.72	Lump or mass in breast	B013497100	Utrogestan capsules	33005B	Mammography
2	217	Benign neoplasm of breast	MA1	Oral liquid preparation	25004C	Surgical pathology Level IV
3	V22.1	Supervision of other normal pregnancy	A037196100	Anpo tablets 10 mg	19014C	Breast ultrasound

Most predictive clinical features for patient classification by cohort. Shows top three features per data modality (ICD-9 diagnostic codes, ATC medication codes, medical orders) ranked by discriminative strength. Complete feature rankings with discriminative weights are provided in online supplemental table 1.

A comparison with existing techniques further contextualises the performance of the proposed MCNN framework. Traditional machine learning models such as Logistic Regression, Decision Trees and Random Forests rely primarily on static feature representations and therefore have limited ability to capture the temporal evolution of clinical records. Gradient boosting methods (eg, XGBoost and LightGBM) achieve stronger performance but still treat input features as aggregated summaries rather than structured temporal sequences. More recent deep learning approaches, including conventional CNNs and transformer-based architectures such as ConvBERT, offer improved feature expressiveness yet operate on single-modality inputs or require substantially larger parameterisation. In contrast, the MCNN integrates multiple clinical modalities through parallel channels and directly models month-to-month patterns over the 36-month observation window, enabling the extraction of both local temporal signals and higher-level trends. This architectural design helps explain the balanced performance observed across cancer types and the model's ability to identify clinically meaningful pre-diagnostic trajectories.

Several limitations require acknowledgement. The reliance on Taiwan NHIRD data may limit generalisability to populations with different healthcare structures and demographics.^{7 26} The absence of lifestyle factors including smoking history, family history and environmental exposures reflects inherent limitations of claims-based databases, as these variables are not systematically recorded in NHIRD rather than being available but

unextracted.²⁷ Additionally, the feature selection procedure was conducted on the complete dataset before cross-validation, which may introduce optimistic bias in performance estimates. Although the selected features represent clinically established cancer-related codes rather than spurious statistical associations, future validation studies should implement nested cross-validation with feature selection performed independently within each outer fold to provide more rigorous generalisation estimates. Furthermore, the use of first cancer-related hospitalisation as the diagnostic index date, while providing standardised and verifiable anchoring, may not capture the true onset of clinical recognition for all patients. Some cancers may have been initially diagnosed through outpatient encounters prior to hospitalisation, meaning the observation window could include early diagnostic or treatment phases rather than exclusively pre-diagnostic periods. This consideration aligns with our SHAP analysis findings, which indicate the model captures patterns during active clinical evaluation rather than purely asymptomatic disease detection. From a methodological standpoint, the class imbalance inherent in population-based cancer detection (3% prevalence) was addressed through stratified cross-validation without additional rebalancing techniques. While this approach evaluates baseline model performance, threshold optimisation during clinical implementation could adjust the sensitivity–specificity trade-off according to specific screening objectives and population risk profiles. Additionally, the fixed 36-month observation window, while supported by prior literature on pre-diagnostic healthcare patterns, may not be optimal

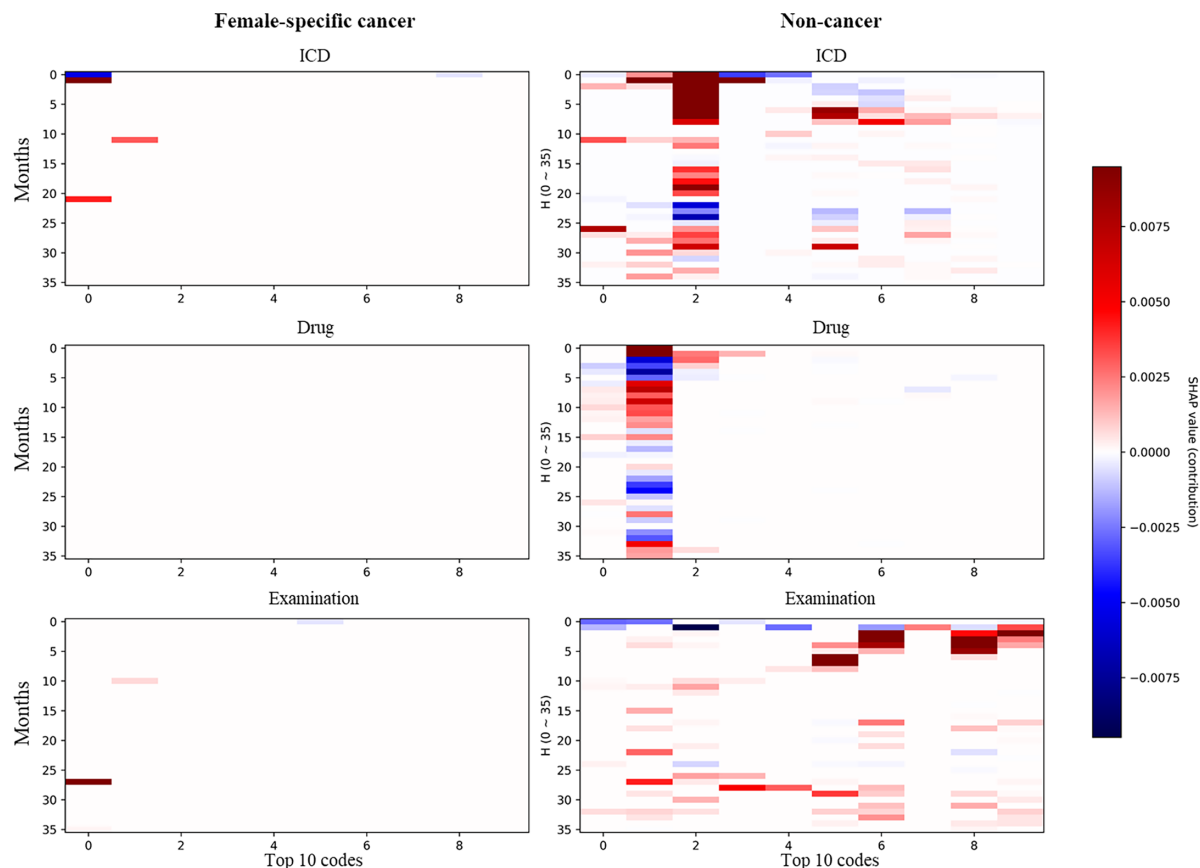


Figure 2 SHAP analysis of temporal feature contributions for female-specific cancer prediction. SHAP values for a representative female-specific patient with cancer across six data channels (diagnostic codes, medications, medical orders×2 classes) during 36-month pre-diagnostic period. Red indicates positive contributions towards female-specific cancer classification; blue indicates negative contributions.

for all cancer types. Future research should examine sensitivity of model performance to varying observation windows. Validation studies using diverse international datasets will be essential for establishing broader applicability. Given similarities in healthcare system structures, findings may be particularly transferable to other East Asian populations with universal coverage and claims-based data systems, though cross-population validation would be required. The integration potential with existing screening programmes offers promising clinical pathways. Rather than replacing current screening modalities, our approach could enhance screening efficiency by identifying high-risk individuals who would benefit most from targeted imaging studies or more frequent surveillance.²⁸ The EHR integration capabilities enable seamless risk assessment during routine clinical encounters without additional data collection burden. Future research should focus on external validation across diverse healthcare systems, incorporation of genetic and lifestyle risk factors and development of clinical decision support interfaces tailored to women's health screening programmes.²⁹ The demonstrated feasibility of leveraging routine healthcare data for accurate cancer risk prediction represents a significant step towards more accessible early detection strategies for female-specific malignancies.

CONCLUSION

This investigation establishes the viability of unified computational detection for breast, ovarian, cervical and uterine cancers through analysis of routine healthcare administrative data. The MCNN achieved Macro-F₁ performance of 0.8443 while maintaining precision of 0.9135, indicating balanced classification capabilities across multiple female malignancy types within a single predictive framework. The 36-month longitudinal approach successfully identified healthcare utilisation patterns that emerge during pre-diagnostic phases, extending detection capabilities beyond current screening programme limitations. The systematic integration of diagnostic codes, medication records and medical orders through multichannel architecture demonstrated complementary predictive value across data modalities. Feature analysis validated clinical relevance through identification of established cancer care pathways, including hormonal therapies and pathological evaluation procedures. The SHAP interpretability framework revealed temporal healthcare engagement patterns that distinguish cancer progression from routine medical care, providing mechanistic insights into administrative data signatures that precede formal clinical recognition.

Computational efficiency achievements through systematic feature selection enable practical healthcare deployment, reducing processing requirements by over 99% while preserving predictive accuracy. The approach addresses current gaps in ovarian and endometrial cancer detection while offering complementary risk stratification for breast and cervical cancer programmes. The moderate specificity performance and Taiwan-specific dataset represent implementation considerations requiring external validation. Future directions include incorporation of lifestyle and genetic risk factors, international dataset validation and clinical decision support interface design tailored to women's health screening workflows.

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 of Taiwan, grant number NSTC 113-2221-E-038-019-MY3 and NSTC 114-2410-H-038-034-MY3.

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 fully anonymised data usage.

Provenance and peer review Part of a topic collection; not commissioned; externally peer reviewed.

Data availability statement Data are available 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

- Ministry of Health and Welfare T. Cancer registry annual report 2024. 2025. Available: <https://dep.moh.gov.tw/DOS/lp-5069-113-xCat-y113.html>
- Nelson HD, Fu R, Cantor A, *et al.* Effectiveness of Breast Cancer Screening: Systematic Review and Meta-analysis to Update the 2009 U.S. Preventive Services Task Force Recommendation. *Ann Intern Med* 2016;164:244–55.
- Saslow D, Solomon D, Lawson HW, *et al.* American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *Am J Clin Pathol* 2012;137:516–42.
- Henderson JT, Webber EM, Sawaya GF. Screening for Ovarian Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA* 2018;319:595–606.
- Siegel RL, Miller KD, Wagle NS, *et al.* Cancer statistics, 2023. *CA A Cancer J Clinicians* 2023;73:17–48.
- Menon U, Gentry-Maharaj A, Burnell M, *et al.* Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): a randomised controlled trial. *The Lancet* 2021;397:2182–93.
- 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.
- Hong M-K, Ding D-C. Early Diagnosis of Ovarian Cancer: A Comprehensive Review of the Advances, Challenges, and Future Directions. *Diagnostics (Basel)* 2025;15:406.
- Brinker TJ, Hekler A, Utikal JS, *et al.* Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. *J Med Internet Res* 2018;20:e11936.
- McKinney SM, Sieniek M, Godbole V, *et al.* International evaluation of an AI system for breast cancer screening. *Nature New Biol* 2020;577:89–94.
- Mariotto AB, Yabroff KR, Shao Y, *et al.* Projections of the cost of cancer care in the United States: 2010–2020. *J Natl Cancer Inst* 2011;103:117–28.
- 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.
- 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.
- World Health Organization. WHO collaborating centre for drug statistics methodology. atc/ddd index 2024. World Health Organization; 2024. Available: https://atcddd.fhi.no/atc_ddd_index
- Yang Y, Pedersen JO. A comparative study on feature selection in text categorization. Proceedings of the fourteenth international conference on machine learning; 1997
- Choi E, *et al.* Doctor ai: predicting clinical events via recurrent neural networks. Machine learning for healthcare conference; 2016
- Guo H, Tang R, Ye Y, *et al.* DeepFM: a factorization-machine based neural network for ctr prediction. Twenty-Sixth International Joint Conference on Artificial Intelligence; Melbourne, Australia, 2017
- Society IC. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2021: IEEE; 2021
- Devlin J, *et al.* Bert: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers); 2019
- Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 2017;30.
- Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019;380:1347–58.
- Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Relevance of breast cancer hormone receptors and other factors to the efficacy of adjuvant tamoxifen: patient-level meta-analysis of randomised trials. *The Lancet* 2011;378:771–84.
- Hammond MEH, Hayes DF, Dowsett M, *et al.* American Society of Clinical Oncology/College Of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Clin Oncol* 2010;28:2784–95.
- Jensen AB, Moseley PL, Oprea TI, *et al.* Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun* 2014;5:4022.
- Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* 2017;376:2507–9.
- Hsing AW, Ioannidis JPA. Nationwide Population Science: Lessons From the Taiwan National Health Insurance Research Database. *JAMA Intern Med* 2015;175:1527–9.
- Torre LA, Siegel RL, Ward EM, *et al.* Global Cancer Incidence and Mortality Rates and Trends—An Update. *Cancer Epidemiol Biomarkers Prev* 2016;25:16–27.
- Smith RA, Andrews KS, Brooks D, *et al.* Cancer screening in the United States, 2019: A review of current American Cancer Society guidelines and current issues in cancer screening. *CA Cancer J Clin* 2019;69:184–210.
- Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44–56.