scientific reports

OPEN

Multitask learning multimodal network for chronic disease prediction

Hsinhan Tsai[1]✉, Ta-Wei Yang[2], Tien-Yi Wu[2], Ya-Chi Tu[1,3], Cheng-Lung Chen[4] & Cheng-Fu Chou[1,2]✉

Chronic diseases are a critical focus in the management of elderly health. Early disease prediction plays a vital role in achieving disease prevention and reducing the associated burden on individuals and healthcare systems. Traditionally, separate models were required to predict different diseases, a process that demanded significant time and computational resources. In this research, we utilized a nationwide dataset and proposed a multi-task learning approach combined with a multimodal disease prediction model. By leveraging patients' medical records and personal information as input, the model predicts the risks of diabetes mellitus, heart disease, stroke, and hypertension simultaneously. This approach addresses the limitations of traditional methods by capturing the correlations between these diseases while maintaining strong predictive performance, even with a reduced number of features. Furthermore, our analysis of attention scores identified risk factors that align with previous research, enhancing the model's interpretability and demonstrating its potential for real-world applications.

Keywords ICD code embedding, Chronic disease, Disease incidence prediction, Multi-task learning, Multimodal network

With advancements in medical standards and lifestyle changes, the global population is aging rapidly. In Taiwan, chronic diseases have become a significant challenge in elderly healthcare. This study examines the challenges posed by chronic diseases from both the patient's and insurer's perspectives.

From the patient's perspective, chronic diseases can lead to severe health complications. Research[1] indicates that older individuals and those with pre-existing conditions such as heart disease, diabetes, or respiratory illnesses are more vulnerable to severe outcomes when facing new health threats like viral infections. Additionally, chronic diseases impose substantial economic burdens on both individuals and society. Early prediction can help prevent or mitigate the impact of these diseases, making it a crucial aspect of healthcare management.

From the insurer's perspective, many individuals purchase insurance to mitigate the financial risks associated with chronic illnesses. However, accurately assessing the health risks of insured individuals is time- and resource-intensive, impacting the efficiency of insurance processes. Predicting the likelihood of future diseases is therefore essential for optimizing risk assessment and resource allocation.

In Taiwan, diabetes mellitus, heart disease, stroke, and hypertension are among the most prevalent chronic diseases, with complex and interrelated causes. Medical literature[2–9] categorizes risk factors into non-modifiable elements—such as age, gender, geographical region, and genetics—and modifiable factors, including hyperlipidemia, metabolic syndrome, and other related conditions. In recent years, several studies have proposed chronic disease prediction models using machine learning (ML) and deep learning (DL)[10–13]. These studies primarily focus on predicting a single chronic disease and have demonstrated good accuracy. However, existing models often fail to consider the interrelationships among multiple chronic diseases and require substantial training time and computational resources. Given the strong correlations among chronic conditions—such as the finding that nearly 25% of individuals aged 14 or older in Madrid suffer from multiple chronic diseases[14], with multimorbidity prevalence reaching up to 30% in Spain[15]—there is significant potential for improvement.

This paper introduces a multi-task learning (MTL) framework for simultaneously predicting multiple chronic diseases by leveraging patients' medical records. Unlike previous studies that rely on lifestyle and biomedical profiles[16,17], our approach focuses on exploring the interactions between diseases while considering temporal

[1]Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. [2]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan. [3]Department of Laboratory Medicine, Linkou Main Branch, Chang Gung Memorial Hospital, Taoyuan, Taiwan. [4]Taiwan Space Agency, Hsinchu, Taiwan. ✉email: hhtsai@cmlab.csie.ntu.edu.tw; ccf@csie.ntu.edu.tw

information, which is crucial for identifying modifiable risk factors and comorbidities. A nationwide dataset from Taiwan validates the robustness of the proposed method. Additionally, attention score analysis highlights key risk factors and comorbidities, aligning with findings in the literature[2–7]. These results underscore the strong interconnections and shared risk factors among diabetes mellitus, heart disease, stroke, and hypertension, emphasizing the efficacy of the MTL model.

Related work

In our study, we build upon the multimodal attention network for dementia (MAND)[18] and extend it to simultaneously predict the risks of diabetes mellitus, heart disease, stroke, and hypertension. We focus on capturing temporal interactions and exploring comorbidities using patients' medical records. To enhance disease representation, we convert International Classification of Diseases (ICD) codes into embeddings using a Word2Vec-based ICD embedding layer. We then employ modules such as multi-head self-attention (MHSA) to capture interactions among these diseases. Furthermore, we investigate comorbidities by applying MTL to predict all four diseases concurrently. Additionally, predicting disease risk from medical records shares similarities with the click-through rate (CTR) prediction problem. Therefore, we also explore the feasibility of both single-task and multi-task CTR models for chronic disease prediction. In this section, we review studies on chronic disease prediction and introduce CTR prediction, MTL techniques, and the MAND framework.

Chronic disease prediction

Several studies have applied ML and DL techniques to chronic disease prediction. For example, ensemble learning has been leveraged to combine the strengths of both ML and DL models[10–12], effectively capturing features and interactions within datasets containing demographic and laboratory information to predict cardiovascular diseases. Similarly, ensemble learning is applied by integrating multiple ML models with feature selection to predict diabetes mellitus using a dataset with various independent medical variables[13]. However, these studies primarily focus on single-disease prediction and do not account for comorbidities among chronic diseases.

To better understand the relationships between chronic diseases, recent studies have increasingly focused on developing models capable of predicting multiple conditions simultaneously using MTL. For instance, an MTL approach has been proposed to jointly predict diabetes and hypertension based on lifestyle, biochemical profiles, and environmental factorsV. Another study introduced an MTL-based method for simultaneously predicting hypertension and diabetes[17], where electronic medical records are first converted into embeddings. The shared convolutional layers capture common information, while task-specific convolutional layers extract disease-specific features. These studies demonstrate that MTL models can achieve prediction performance comparable to single-task learning (STL) models optimized for specific diseases. In other words, the strong correlation between hypertension and diabetes suggests that jointly modeling these diseases may lead to more comprehensive and effective prediction models. However, these studies do not account for how past medical conditions and temporal interactions influence the development of chronic diseases.

Click through rate

Click-through rate (CTR) prediction is a widely studied topic in recommendation systems. CTR models analyze customer data such as browsing and purchase histories to predict the likelihood of a customer clicking on advertisements in the future. Factorization machines (FMs)[19] have been particularly effective in capturing feature interactions, paving the way for FM-based models like DeepFM[20] and xDeepFM[21]. Beyond FM-based approaches, neural network (NN)-based models such as deep and cross network (DCN)[22] and its extension DCNV2[23] have been developed to tackle CTR prediction, utilizing deep learning to capture complex patterns and enhance prediction accuracy.

This paper takes inspiration from the core concept of CTR prediction: learning interactions between features. In our context, patient medical records are analogous to customer browsing histories. By learning the interactions between diseases, the model predicts the risk of future disease occurrences in patients. This approach has already shown impressive results in predicting dementia[18].

Multi-task learning

The objective of multi-task learning is to simultaneously learn multiple tasks, leveraging shared representations to enhance generalization across tasks. Common approaches to implementing MTL are categorized into the hard parameter sharing and the soft parameter sharing[24].

- Hard parameter sharing: This method shares hidden layers among tasks, enabling the learning of shared representations. The weights of the shared hidden layers are updated based on the influence of all tasks, which significantly reduces the risk of overfitting[25].
- Soft parameter sharing: In this approach, each task has its own independent model and parameters. Unlike single-task learning, constraints are applied to the differences between the parameters of each model during training. Techniques such as l_2 norm (Euclidean distance)[25] or trace norm[26] are commonly used to enforce these constraints.

While MTL can mitigate overfitting, it is susceptible to negative transfer, particularly when tasks have weak or conflicting relationships. This issue, known as the seesaw phenomenon[27], refers to the counterproductive effect where improvements in one task lead to a decline in performance in another.

In this study, we treat the risk prediction of diabetes mellitus, heart disease, stroke, and hypertension as distinct tasks. Fortunately, these tasks exhibit strong interrelationships. For example, diabetes increases the risk of heart disease and stroke[3–6]; hypertension and stroke are closely related[5,6]; and diabetes patients are prone to complications such as heart disease and stroke[4,5]. These correlations indicate that our tasks are well-aligned, preventing the occurrence of weak task relationships and the seesaw phenomenon.

Multimodal attention network for dementia

The multimodal attention network for dementia (MAND) framework predicts a patient's likelihood of developing dementia within 5 years using medical records and personal information from the past 10 years[18]. The medical records include a history of diseases represented by ICD codes. MAND selects the five most frequently occurring diseases every 2 months over 10 years, generating 300 disease-related features that capture both disease history and temporal patterns. The ICD codes are then transformed into embeddings using a pre-trained Word2Vec-based ICD embedding layer. To model interactions among these embeddings, MAND incorporates multiple feature extraction modules, including MLP, LSTM, CNN, and multi-head self-attention (MHSA). The model achieves strong predictive performance and identifies potential risk factors through attention score analysis.

In our work, we adapt the MAND framework to predict diabetes mellitus, heart disease, stroke, and hypertension. Additionally, we extend MAND to an MTL framework that predicts these four chronic diseases simultaneously. Our attention score analysis is also expanded to investigate not only key risk factors but also disease comorbidities.

Dataset and methodology

In this section, we present the problem definition for disease prediction, describe the dataset, and outline the proposed architecture of the disease prediction model.

Problem definition

We leverage a patient's medical records from the past 10 years, along with personal information, to predict their risk of developing diabetes mellitus, heart disease, stroke, or hypertension within the next 5 years. This problem is similar to that described in the literature[18], which focuses on dementia prediction.

Dataset

The data were retrospectively collected from the Health and Welfare Data Science Center (HWDC) in Taiwan, which provides randomly sampled medical records from two million individuals for academic research purposes. The study was approved by the Research Ethics Committee of National Taiwan University (NTU-REC No.: 202108HM002) on August 9, 2021. Informed consent was waived by the Research Ethics Committee of National Taiwan University due to the retrospective nature of the study. Additionally, the study was conducted in accordance with the ethical guidelines outlined in the Declaration of Helsinki.

Our study focuses on predicting the 5-year occurrence of diabetes mellitus, heart disease, stroke, and hypertension using personal information and medical records from the past 10 years. We extract patients' personal information from the "Registry for Beneficiaries" file and medical records from the "Ambulatory Care Expenditures" file. Patient age is calculated from the ID_BIRTH_Y column, whereas medical records are identified using the ICD_CM column, which contains ICD codes for diagnosed conditions. The corresponding visit time is derived from the FEE_YM column.

To construct disease features, we select the three most frequently occurring diseases every 2 months over the past 10 years, resulting in a feature length of $3 \times 6 \times 10 = 180$. Missing values are replaced with a <PAD> token. We exclude patients who were diagnosed with any of the target diseases (diabetes mellitus, heart disease, stroke, or hypertension) during the first 10 years. The final dataset consists of 555,124 samples. The occurrence rates for diabetes mellitus, heart disease, stroke, and hypertension are 22.4%, 24.6%, 8.7%, and 39.0%, respectively. The proportion of negative cases (patients who did not develop any of these four diseases) is 51.3%.

Registry for beneficiaries

The personal information fields stored in this database are described as follows:

- ID: The unique identification code for each individual, serving as the primary key for database merging.
- ID_BIRTH_Y: The patient's year of birth, also utilized as a key for database merging.
- ID_S: The patient's gender.
- HOME_CITY: The residential area code of the patient, providing detailed regional information.
- REMOTE_MARK: An indicator of whether the patient resides in a remote area, capturing broader regional information.
- ID_IDENT: The category of the patient's occupation.

Ambulatory care expenditures

The medical record fields stored in this database are described as follows:

- ID: The unique identification code for each individual, serving as the primary key for database merging.
- AGE: The patient's age at the time of medical treatment, used both as a key for database merging and as part of the model's input.
- FEE_YM: The year and month of the patient's medical treatment.
- ICD_CM: The diagnosis code associated with the patient's current medical visit.

Model architecture

In this research, we utilize three models to achieve the goal of disease prediction. The first is the pre-trained ICD Word2Vec model, which is employed for ICD embedding. The second is the single-task learning (STL) disease prediction model, which is designed for predicting a specific disease. The third and most significant is the multi-task learning (MTL) disease prediction model, which is proposed in this study as the primary innovation.

Pre-trained ICD Word2Vec Model
This model is designed to transform ICD codes into embeddings in a latent space, grouping the embeddings of similar diseases for improved representation. In our dataset, all ICD codes diagnosed for a patient within a 2-month interval are treated as a sentence. ICD codes appearing in the same interval are considered positive examples; those from different intervals are treated as negative examples. To learn meaningful ICD embeddings efficiently, we apply the skip-gram model[28] to the ICD corpus and use negative sampling[29] to reduce computational complexity. These techniques allow the model to automatically generate high-quality ICD embeddings while minimizing parameter complexity.

Single-task learning disease prediction model
This paper employs six STL models for disease prediction, consisting of four multimodal models with different ICD extraction modules and two CTR models. The multimodal model illustrated in Fig. 1a is referred to as MAND[18]. This framework provides flexibility in selecting the ICD extraction module for various purposes. We define MAND-LR, MAND-MLP, MAND-LSTM, and MAND-MHSA to represent the MAND architecture integrated with LR, MLP, LSTM, and MHSA as ICD extraction modules, respectively.

The FM[19] and DCN[22] CTR models serve as representatives of FM-based and NN-based approaches, respectively. Model inputs are divided into three categories: medical records represented by ICD code sequences, numerical features derived from personal information, and categorical features of personal information.

Multi-task learning disease prediction model
We expanded the single-disease prediction model into an MTL model. The MTL multimodal network architecture, illustrated in Fig. 1b, employs hard parameter sharing, with all parameters above the concatenation layer being shared. This approach significantly reduces the number of parameters while enabling the latent representations learned by the ICD extraction module to simultaneously capture the relationships between diabetes mellitus, heart disease, stroke, and hypertension.

Extending the CTR model to support MTL is relatively straightforward. This is achieved by separately sharing the FM layer outputs from Fig. 2a and the concatenation outputs from Fig. 2b to construct an MTL CTR model.

Results

In our experiments, we split the dataset of 555,124 samples into training, validation, and testing sets with a ratio of 0.64, 0.16, and 0.2, respectively. The construction of disease features is described in the Dataset section. Each sample contains 180 ICD codes, with every three ICD codes representing a 2-month interval.

Data preprocessing includes standardization of numerical features, label encoding for categorical features, and tokenization of ICD codes. For pretraining the ICD embedding layer, we use cosine similarity to measure the distance between ICD embeddings in the feature space.

To train the MAND prediction model (illustrated in Fig. 1a), we compared different ICD extraction modules, including LR, MLP, LSTM, and MHSA. The hyperparameters for each model were set as follows: the MLP module consisted of two hidden layers with 256 and 128 neurons. For MHSA, the number of heads, key dimension, and value dimension were set to 1, 16, and 8, respectively. For the FM model (illustrated in Fig. 2a), the FM

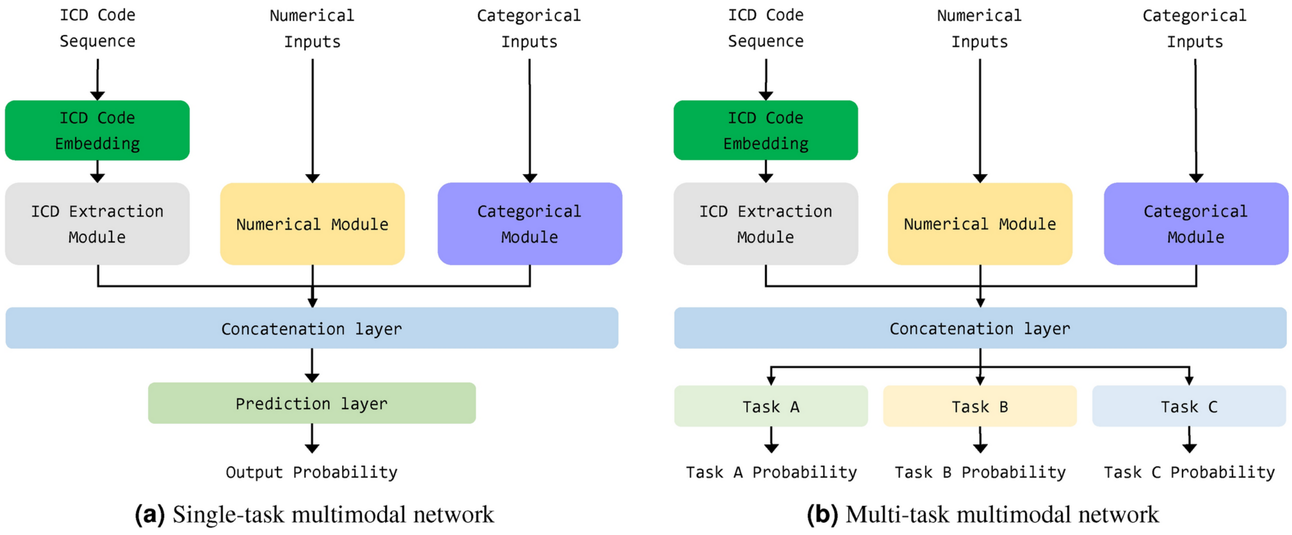
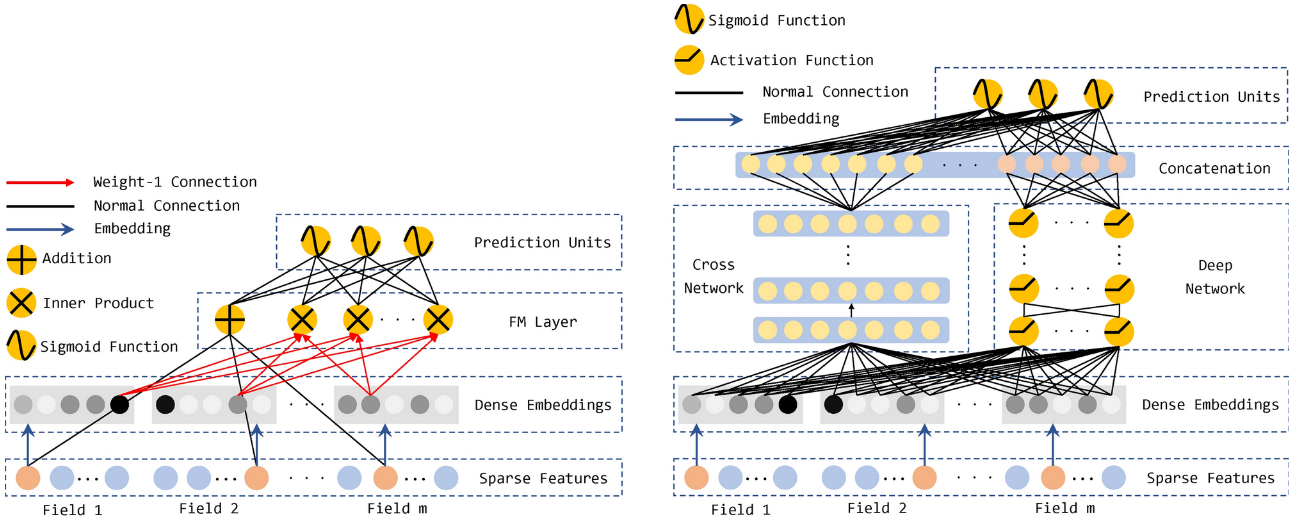


Fig. 1. The architectures of (**a**) the single-task model[18] and (**b**) the multi-task multimodal network are shown. The ICD extraction module can be implemented using logistic regression, MLP, LSTM, or multi-head self-attention.



(a) Multi-task factorization machine

(b) Multi-task deep & cross network

Fig. 2. Two CTR models of multi-task architectures: (**a**) multi-task factorization machine (**b**) multi-task deep and cross network.

layer had a dimensionality of 8. In the DCN model (illustrated in Fig. 2b), the two hidden layers contained 32 neurons each, and the ReLU activation function was applied. We used the Adam optimizer and log loss as the loss function. Early stopping was applied by monitoring the AUC value on the validation set, stopping training if no improvement was observed within five epochs. All results were evaluated on the testing set.

In this section, we first demonstrate the effectiveness of Word2Vec in reducing the dimensionality of high-dimensional ICD codes. Next, we emphasize the comparable performance of our proposed MTL models, encompassing both multimodal networks and CTR models. We also show that MTL models require only slightly more parameters than STL models while enabling shared learning across tasks. Finally, we investigate the interpretability of these models in relation to medical records and personal information. The performance of the multi-task learning models is evaluated using metrics such as log loss, AUC, balanced accuracy (BAC), F1 score, false positive rate (FPR), and false negative rate (FNR), ensuring a comprehensive assessment of their capabilities.

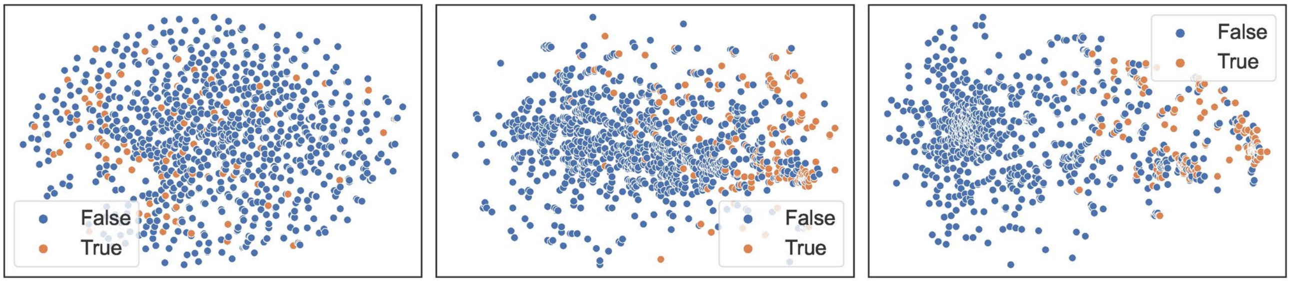
Data visualization

Traditionally, one-hot encoding is used for categorical representation; however, due to the large number of ICD codes, this can lead to memory constraints. For example, if the total number of ICD codes is *n*, the one-hot encoded vector $x_{\text{one}} \in \{0,1\}^n$ has a dimension of *n*. To mitigate this issue, we apply label encoding and multi-hot encoding as alternatives. In label encoding, each ICD code is assigned a unique integer value, resulting in the representation $x_{\text{lab}} \in \mathbb{Z}_0^+$ with a dimension of 1. Meanwhile, multi-hot encoding represents each ICD code using a binary vector of size $\lceil \log_2 n \rceil$, denoted as $x_{\text{mul}} \in \{0,1\}^{\lceil \log_2 n \rceil}$. Both encoding methods significantly reduce the dimensionality compared to one-hot encoding. Finally, we compare the performance of Word2Vec-based ICD embeddings with label encoding and multi-hot encoding. The Word2Vec embedding is represented as $x_{\text{emb}} \in \mathbb{R}^k$, where $k = 8$ in our experiment. To visualize the distribution of ICD codes, we employ t-SNE (t-distributed stochastic neighbor embedding)[30], a nonlinear dimensionality reduction technique that maps high-dimensional data into a low-dimensional space. In our study, t-SNE represents each ICD code as a two-dimensional point, where similar diseases are positioned closer together, while unrelated diseases are mapped farther apart with high probability.

Figures 3, 4, 5, and 6 present the visualization results for heart disease, diabetes mellitus, stroke, and hypertension, respectively. In these figures, we use “True” to indicate that a sample point corresponds to a patient diagnosed with the disease and “False” to indicate that the patient is not diagnosed with the disease. Taking heart disease as an example, Fig. 3a demonstrates that label encoding fails to distinguish between patients with and without heart disease. In contrast, Fig. 3b shows significant improvement with multi-hot encoding, where patients with heart disease (orange) are more concentrated on the right side. Moreover, Fig. 3c highlights the superior clustering effect of the pre-trained ICD embedding—patients without heart disease (blue) form a large, distinct cluster, whereas those with the disease (orange) form another clearly defined cluster. Similar patterns can be observed for the other three diseases. These results confirm that the ICD embedding effectively captures relationships between different ICD codes, providing a robust representation of diseases.

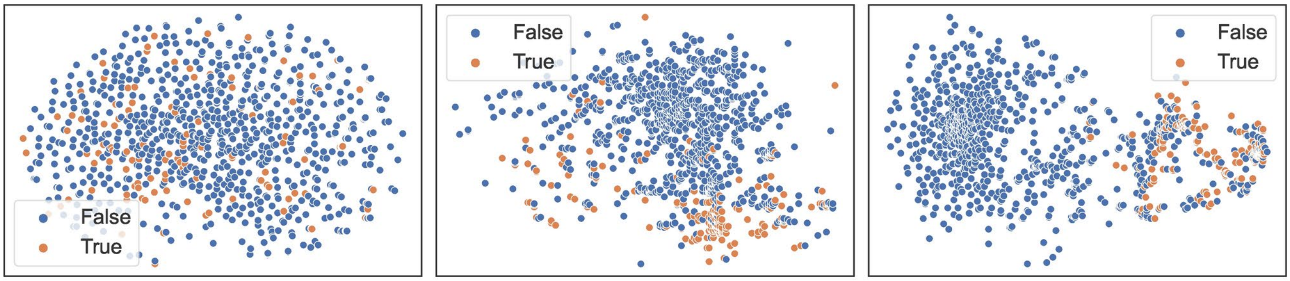
Performance: single-task versus multi-task learning

In this experiment, we directly compare the performance of six models under both STL and MTL settings. Tables 1, 2, 3, and 4 present the results for heart disease, diabetes mellitus, stroke, and hypertension, respectively. Overall, we observe that both STL and MTL approaches achieve high AUC scores across all four diseases. First,



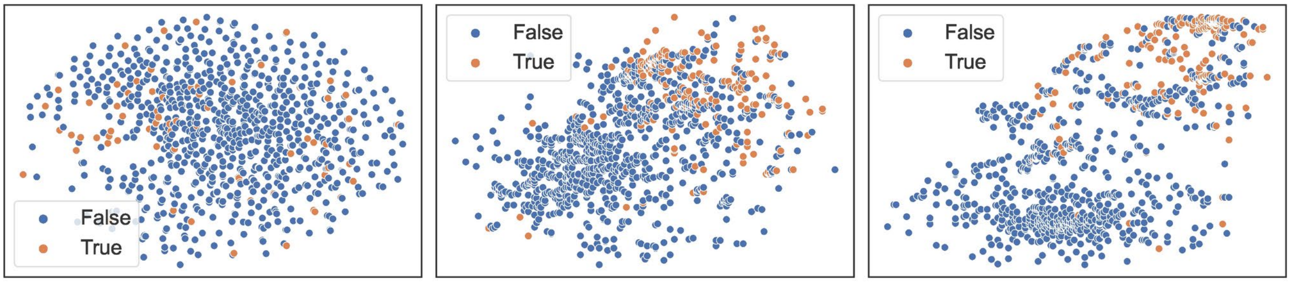
(a) Label encoding **(b)** Multi-hot encoding **(c)** Word2Vec ICD embedding

Fig. 3. Heart disease data visualization in different encodings.



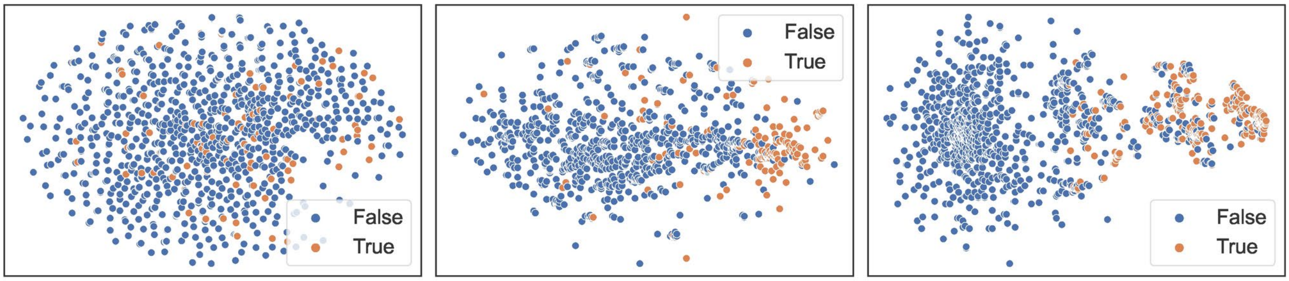
(a) Label encoding **(b)** Multi-hot encoding **(c)** Word2Vec ICD embedding

Fig. 4. Diabetes mellitus data visualization in different encodings.



(a) Label encoding **(b)** Multi-hot encoding **(c)** Word2Vec ICD embedding

Fig. 5. Stroke data visualization in different encodings.



(a) Label encoding **(b)** Multi-hot encoding **(c)** Word2Vec ICD embedding

Fig. 6. Hypertension data visualization in different encodings.

Backbone model	STL/MTL	Log loss	AUC	BAC	Precision	Recall	F1 score	FPR	FNR
MAND-LR[18]	STL	**0.3838**	**0.8602**	**0.7067**	**0.7422**	**0.4663**	**0.5728**	**0.0529**	**0.5337**
	MTL	0.3861	0.8580	0.7002	0.7352	0.4535	0.5609	0.0531	0.5465
MAND-MLP[18]	STL	**0.3581**	**0.8698**	0.7204	**0.8041**	0.4789	0.6003	**0.0381**	0.5211
	MTL	0.3640	0.8695	**0.7239**	0.7507	**0.5019**	**0.6016**	0.0541	**0.4981**
MAND-LSTM[18]	STL	0.3543	0.8774	0.7162	**0.8555**	0.4576	0.5962	**0.0252**	0.5424
	MTL	**0.3467**	**0.8787**	**0.7205**	0.8366	**0.4710**	**0.6026**	0.0300	**0.5290**
MAND-MHSA[18]	STL	0.3621	**0.8765**	**0.7512**	0.7224	**0.5744**	**0.6400**	0.0720	**0.4256**
	MTL	**0.3614**	0.8579	0.7411	**0.7286**	0.5486	0.6259	**0.0664**	0.4514
FM[19]	STL	0.4421	0.8467	0.7012	0.7056	0.4658	0.5612	**0.0634**	0.5342
	MTL	**0.3915**	**0.8576**	**0.7215**	**0.7084**	**0.5114**	**0.5940**	0.0684	**0.4886**
DCN[22]	STL	0.3566	0.8745	0.7232	**0.8123**	0.4829	0.6057	**0.0365**	0.5171
	MTL	**0.3523**	**0.8749**	**0.7308**	0.7867	**0.5062**	**0.6160**	0.0446	**0.4938**

Table 1. Comparison between STL and MTL in heart disease prediction. MAND-LR, MAND-MLP, MAND-LSTM, and MAND-MHSA denote the MAND architecture integrated with logistic regression, multilayer perceptron (MLP), LSTM, and multi-head self-attention as ICD feature extraction modules, respectively. FM and DCN represent CTR-based approaches. BAC: balanced accuracy; FPR: false positive rate; FNR: false negative rate. Bold font indicates the better performance values between STL and MTL.

Backbone model	STL/MTL	Log loss	AUC	BAC	Precision	Recall	F1 score	FPR	FNR
MAND-LR[18]	STL	**0.3043**	**0.8818**	**0.7588**	0.9331	**0.5284**	**0.6748**	0.0108	**0.4716**
	MTL	0.3084	0.8803	0.7563	**0.9337**	0.5233	0.6707	**0.0107**	0.4767
MAND-MLP[18]	STL	**0.2906**	**0.8852**	**0.7665**	**0.9638**	**0.5387**	**0.6911**	**0.0057**	**0.4613**
	MTL	0.2967	0.8831	0.7607	0.9598	0.5277	0.6810	0.0063	0.4723
MAND-LSTM[18]	STL	0.2858	**0.8926**	**0.7791**	0.9436	**0.5680**	**0.7091**	0.0098	**0.4320**
	MTL	**0.2850**	0.8912	0.7728	**0.9543**	0.5533	0.7005	**0.0077**	0.4467
MAND-MHSA[18]	STL	**0.2888**	**0.8918**	**0.7822**	0.9003	**0.5829**	**0.7076**	0.0185	**0.4171**
	MTL	0.2924	0.8900	0.7760	**0.9019**	0.5698	0.6984	**0.0178**	0.4302
FM[19]	STL	0.3378	**0.8749**	**0.7635**	0.8557	**0.5538**	**0.6725**	0.0268	**0.4462**
	MTL	**0.3250**	0.8699	0.7551	**0.8761**	0.5319	0.6620	**0.0217**	0.4681
DCN[22]	STL	**0.2871**	**0.8899**	**0.7749**	0.9493	**0.5584**	**0.7031**	0.0086	**0.4416**
	MTL	0.2914	0.8870	0.7690	**0.9495**	0.5464	0.6936	**0.0084**	0.4536

Table 2. Comparison between STL and MTL in diabetes mellitus prediction. MAND-LR, MAND-MLP, MAND-LSTM, and MAND-MHSA denote the MAND architecture integrated with logistic regression, multilayer perceptron (MLP), LSTM, and multi-head self-attention as ICD feature extraction modules, respectively. FM and DCN represent CTR-based approaches. BAC: balanced accuracy; FPR: false positive rate; FNR: false negative rate. Bold font indicates the better performance values between STL and MTL.

this demonstrates that the multimodal architecture[18] developed for dementia prediction can also effectively predict chronic diseases. Additionally, CTR models originally designed for recommendation systems can be adapted for disease prediction. Second, there is a common assumption that STL sets an upper bound for MTL performance. However, we observe that some MTL results outperform STL results, suggesting that shared information or multimorbidity may exist among these four chronic diseases. Third, regardless of STL or MTL, the best AUC results are achieved by MAND-LSTM, highlighting the importance of temporal information extracted from ICD code sequences.

However, we observe high FNRs in heart disease, diabetes mellitus, and stroke prediction, indicating that the models tend to predict negative cases more frequently. This issue may stem from data imbalance, as the prevalence rates of heart disease, diabetes mellitus, and stroke in the dataset are 24.6%, 22.4%, and 8.7%, respectively. The highest FNR occurs in stroke prediction, which corresponds to the lowest occurrence rate, further highlighting the impact of data imbalance.

Number of parameters: STL versus MTL

Another advantage of the MTL approach is its ability to predict multiple diseases simultaneously while requiring only a small number of additional parameters compared to STL models. This is because MTL models share a significant number of weights across tasks. To demonstrate this, we compare the number of parameters required for STL and MTL in Table 5. Since our study predicts four diseases, we also present the total number of parameters required for STL when predicting all four conditions separately. From this table, we observe that although MTL

Backbone model	STL/MTL	Log loss	AUC	BAC	Precision	Recall	F1 score	FPR	FNR
MAND-LR[18]	STL	**0.2120**	0.8585	**0.6023**	0.7794	**0.2102**	**0.3311**	0.0056	**0.7898**
	MTL	0.2134	**0.8586**	0.6009	**0.7857**	0.2071	0.3279	**0.0053**	0.7929
MAND-MLP[18]	STL	**0.2042**	**0.8626**	**0.6169**	**0.8939**	**0.2364**	**0.3739**	0.0026	**0.7636**
	MTL	0.2198	0.8467	0.5694	0.8637	0.1409	0.2422	**0.0021**	0.8591
MAND-LSTM[18]	STL	**0.1974**	**0.8700**	**0.6460**	**0.8339**	**0.2977**	**0.4387**	0.0057	**0.7023**
	MTL	0.2050	0.8625	0.6286	0.8240	0.2626	0.3983	**0.0054**	0.7374
MAND-MHSA[18]	STL	**0.2020**	**0.8703**	**0.6445**	**0.7540**	**0.2982**	**0.4274**	**0.0092**	**0.7018**
	MTL	0.2076	0.8643	0.6392	0.7367	0.2882	0.4143	0.0098	0.7118
FM[19]	STL	0.2619	0.8330	**0.6226**	0.5708	**0.2642**	0.3612	0.0190	**0.7358**
	MTL	**0.2288**	**0.8351**	0.6216	**0.6718**	0.2551	**0.3698**	**0.0119**	0.7449
DCN[22]	STL	**0.2013**	**0.8649**	**0.6336**	**0.8600**	**0.2715**	**0.4127**	**0.0043**	**0.7285**
	MTL	0.2040	0.8627	0.6288	0.8509	0.2620	0.4003	0.0044	0.7380

Table 3. Comparison between STL and MTL in stroke prediction. MAND-LR, MAND-MLP, MAND-LSTM, and MAND-MHSA denote the MAND architecture integrated with logistic regression, multilayer perceptron (MLP), LSTM, and multi-head self-attention as ICD feature extraction modules, respectively. FM and DCN represent CTR-based approaches. BAC: balanced accuracy; FPR: false positive rate; FNR: false negative rate. Bold font indicates the better performance values between STL and MTL.

Backbone model	STL/MTL	Log loss	AUC	BAC	Precision	Recall	F1 score	FPR	FNR
MAND-LR[18]	STL	**0.3383**	0.9232	**0.8324**	0.8887	**0.7227**	**0.7972**	0.0579	**0.2773**
	MTL	**0.3383**	**0.9234**	0.8320	**0.8980**	0.7158	0.7966	**0.0518**	0.2842
MAND-MLP[18]	STL	**0.3073**	**0.9300**	**0.8450**	**0.9002**	0.7426	**0.8138**	**0.0526**	0.2574
	MTL	0.3127	0.9276	0.8427	0.8918	**0.7429**	0.8105	0.0575	**0.2571**
MAND-LSTM[18]	STL	0.2990	**0.9346**	0.8506	**0.9213**	0.7416	0.8217	**0.0404**	0.2584
	MTL	**0.2949**	**0.9346**	**0.8519**	0.9179	**0.7463**	**0.8233**	0.0425	**0.2537**
MAND-MHSA[18]	STL	**0.3069**	0.9325	0.8539	**0.8907**	0.7681	0.8249	**0.0603**	0.2319
	MTL	**0.3069**	**0.9331**	**0.8558**	0.8845	**0.7762**	**0.8268**	0.0646	**0.2238**
FM[19]	STL	0.3821	0.9138	0.8236	0.8772	0.7108	0.7853	0.0636	0.2892
	MTL	**0.3394**	**0.9202**	**0.8323**	**0.8971**	**0.7170**	**7970**	**0.0524**	**0.2830**
DCN[22]	STL	**0.3017**	**0.9319**	**0.8494**	0.8946	**0.7557**	**0.8193**	0.0569	**0.2443**
	MTL	0.3058	0.9302	0.8459	**0.9070**	0.7403	0.8152	**0.0485**	0.2597

Table 4. Comparison between STL and MTL in hypertension prediction. MAND-LR, MAND-MLP, MAND-LSTM, and MAND-MHSA denote the MAND architecture integrated with logistic regression, multilayer perceptron (MLP), LSTM, and multi-head self-attention as ICD feature extraction modules, respectively. FM and DCN represent CTR-based approaches. BAC: balanced accuracy; FPR: false positive rate; FNR: false negative rate. Bold font indicates the better performance values between STL and MTL.

	Parameters		
Model	STL	4 STLs	MTL
MAND-LR[18]	178,782	715,128	178,860
MAND-MLP[18]	402,117	1,608,468	419,784
MAND-LSTM[18]	187,333	749,332	195,784
MAND-MHSA[18]	181,445	725,780	187,208
FM[19]	402,345	1,609,380	402,385
DCN[22]	416,441	1,665,764	421,196

Table 5. Comparison of number of parameters between STL and MTL across various models. “4STLs” represents the total parameters required to predict heart disease, diabetes mellitus, stroke, and hypertension using four separate STL models.

Mask rate	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Log loss	0.2829	0.2838	0.2857	0.2898	0.2933	0.2946	0.3009	0.3113	0.3211	0.3845	0.5348
AUC	0.8918	0.8911	0.8900	0.8878	0.8863	0.8856	0.8825	0.8785	0.8752	0.8577	0.8031

Table 6. Performance impact of randomly masking medical records, highlighting their importance in disease prediction.

	Log loss	AUC
No permutation	0.2921	0.8908
Age	**0.3338**	**0.8364**
Gender	0.2931	0.8899
Remote area	0.2922	0.8907
Residential area	0.2937	0.8890
Occupation	0.2924	0.8904

Table 7. Permutation feature importance of personal information. A greater AUC drop compared to "No permutation" indicates higher feature importance. Bold font indicates the largest performance drop.

models require slightly more parameters than a single STL model due to disease-specific prediction layers, the shared common layers may help uncover comorbidities, which is crucial in clinical research. In contrast, STL models require four times the parameters of a single STL model to predict four diseases, making them significantly more demanding in terms of computational and storage resources. This could pose challenges in resource-limited environments. For instance, deploying four separate STL models on a constrained edge device would require substantially more storage, whereas an MTL model would only require about one-fourth of that capacity. Additionally, the overhead of loading multiple STL models could introduce latency issues, making real-time predictions less feasible. These findings highlight the advantages of MTL in terms of both computational efficiency and storage optimization, making it a practical choice for real-world clinical applications such as clinical decision support.

Feature importance

In this experiment, we separately evaluate the importance of medical records and personal information to assess the model's robustness and determine whether its results align with findings from the medical literature.

Medical records

Since the MAND architecture is designed to extract interactions among medical records, we investigated the importance of medical records by randomly masking a proportion thereof. We examined whether the model could maintain its predictive capability when medical records were partially or entirely missing. The ICD extraction module applied in this experiment was LSTM. We masked medical records by replacing them with a<PAD> token, following the approach described in MAND[18]. The performance was computed by averaging the metrics across four tasks.

As shown in Table 6, using the complete medical records achieved an AUC of 0.8918. In contrast, completely masking the medical records (relying solely on personal information) resulted in an AUC of 0.8031, indicating that medical records are crucial for the model. Including complete medical records led to a 9-percentage-point increase in AUC. Further analysis of the results with different masking proportions revealed that when 50% to 60% of the medical records were masked, the AUC decreased by only approximately 1 percentage point, demonstrating its robustness and effective utilization of medical record interactions.

Personal information

In this section, we used permutation feature importance[31] to assess the impact of various personal information features on the model's predictions. The greater the performance drop caused by permuting a feature, the more important that feature is to the model.

As shown in Table 7, age has the highest influence on the model's predictions. This finding aligns with existing research, which identifies age as a critical non-modifiable risk factor for chronic diseases such as diabetes mellitus, heart disease, stroke, and hypertension[2,3,5–8]. As individuals age, the decline in various physiological functions significantly increases their risk of developing these conditions.

Other features do not lead to a drastic decline in model performance. However, we still observe that gender and residential area may be potential factors, consistent with insights from the literature[2,5,6,8,9].

ICD interpretability: high attention score

In this experiment, we leveraged attention scores from the self-attention mechanism to interpret ICD codes. Specifically, we selected the top 2000 patients with the highest prediction scores for diabetes mellitus, heart disease, stroke, and hypertension, as these can be seen as high-confidence model predictions. We then identified ICD code pairs with the highest attention scores among these patients, highlighting the relationships between specific ICD codes that the model deemed most significant. These ICD code pairs were categorized into the following three groups:

- *Modifiable risk factor*. Pure hypercholesterolemia (272.0); Mixed hyperlipidemia (272.2); and Other and unspecified hyperlipidemia (272.4).
- *Multimorbidity*. Gout, unspecified (274.9); Senile cataract, unspecified (366.10); Chronic hepatitis, unspecified (571.40); Chronic renal failure (585); Osteoarthrosis, localized, not specified whether primary or secondary, lower leg (715.36); Osteoarthrosis, unspecified whether generalized or localized, unspecified site (715.90); and Lumbosacral spondylosis without myelopathy (721.3).
- *Emerging factor*. Anxiety state, unspecified (300.00) and Neurotic depression (300.4).

Discussion

In this study, we demonstrate the feasibility of multi-task learning in simultaneously predicting diabetes mellitus, heart disease, stroke, and hypertension. We also extend the capabilities of the multimodal model developed in our previous research[18], which originally focused on dementia prediction. Our findings confirm that the model's effectiveness is not limited to a single disease but can be generalized to multiple conditions, transitioning from single- to multi-task learning.

We highlight the importance of Word2Vec-based ICD embeddings in refining the feature space by grouping related diseases while distancing unrelated ones. Unlike label encoding and multi-hot encoding, which lack semantic meaning, Word2Vec embeddings learn the relationships between diseases, leading to a more separable and meaningful feature representation.

Our results show that MTL models perform comparably to STL models in predicting the four chronic diseases. This improvement may be attributed to the strong correlation among these diseases, as they often serve as risk factors for one another[2–7]. Prior research[14,15] also supports the presence of multimorbidity among chronic diseases. By leveraging MTL, we effectively capture shared features across tasks, leading to improved overall performance. Despite minor performance variations existing due to disease-specific differences, our results emphasize the value of shared features and suggest the potential for uncovering common risk factors.

In our feature importance experiments, we found that even with 60% of ICD codes masked, the model maintained a high AUC of 0.8825. This suggests that a large number of effective features exist or that many redundant diseases appear in ICD code sequences. It also implies that key risk factors are common across different patients. Consequently, even with only a subset of important risk factors available, the model retains strong predictive performance, demonstrating its resilience and adaptability.

Furthermore, by selecting the top 2000 patients with the highest prediction scores, we analyzed disease correlations in cases where the model made high-confidence predictions. This approach helped identify key diseases, some of which are established modifiable risk factors for diabetes, heart disease, stroke, and hypertension[3–7,9]. Others, such as gout, senile cataract, and chronic hepatitis, are chronic conditions commonly found in elderly populations. Extensive medical studies have shown that patients with chronic diseases are more likely to develop multimorbidity[14,15,32], explaining the higher attention scores assigned to these conditions by our model. Additionally, emerging research suggests that anxiety and depression may increase the likelihood of developing chronic diseases[32,33], making them potential risk factors for further investigation. Note that although the balanced accuracy of our model ranges from the 60%s to the 80%s across the four diseases (from Table 1, 2, 3, 4), we believe the results remain meaningful. While a model with accuracy below 90% may not be sufficient for independent diagnosis, medical AI systems are typically designed to support clinical decision-making rather than replace it. Even without reaching 90% accuracy, such models can offer valuable insights or explanations to clinicians. Considering the complexity of chronic diseases and the presence of many hidden risk factors, we believe that a model achieving around 80% accuracy can still help doctors make more informed decisions when used in conjunction with their clinical judgment.

Despite these promising findings, this study has several limitations. First, the models rely on patients' medical records, which are sensitive and may pose challenges in real-world implementation. Second, our experiments highlight age as an important predictor, but further research is needed to identify key risk factors in younger populations for early prevention. Third, we use patients' medical records from the past 10 years to predict disease incidence over the next 5 years. Future studies could explore whether shorter medical histories (e.g., 5 years) predict long-term disease risk (e.g., 10 years). Last, data imbalance may contribute to a high false negative rate. Although we preserved the original data distribution to avoid disrupting comorbidity analysis, it remains valuable to explore techniques that can effectively address imbalance in MTL models without compromising the insights into disease relationships.

Future directions for this research include simplifying model input, conducting age-stratified analysis, extending to long-term disease prediction, addressing data imbalance among diseases, and predicting yearly disease incidence. To simplify input data, a potential approach is model distillation, reducing the required medical records to only key features, making the model more practical for real-world applications, such as clinical decision support. Age-stratified analysis would allow us to explore risk factors across different age groups, providing more personalized insights. Long-term disease prediction could enable earlier intervention for chronic conditions. Additionally, the high false negative rate caused by data imbalance might be mitigated through techniques such as sampling and cost-sensitive learning. Lastly, predicting disease incidence on a yearly basis could help identify risk factors that accelerate disease onset. We leave these directions for future research.

Conclusion

Building upon our previous research, we have proposed multi-task learning models to predict multiple chronic diseases simultaneously. Our findings not only validate the feasibility of applying our previous model to different diseases but also highlight the potential of multi-task learning in this domain. The comparable performance of multi-task learning and single-task learning suggests the presence of common risk factors among these diseases.

Furthermore, the model's ability to maintain strong performance with only 40% of medical records demonstrates its resilience and potential for knowledge distillation in future work. Additionally, diseases with high attention scores align with findings from previous studies, further reinforcing the reliability of our results.

Data availability

The datasets analyzed during the current study are not publicly available because they were used under license from the Health and Welfare Data Science Center in Taiwan for the current study; however, they are available from the corresponding author on reasonable request.

Received: 12 February 2025; Accepted: 21 April 2025
Published online: 03 May 2025

References

1. Pal, R. & Bhadada, S. K. Covid-19 and non-communicable diseases. *Postgrad. Med. J.* **96**, 429–430 (2020).
2. Sheen, Y. J. et al. Trends in prevalence and incidence of diabetes mellitus from 2005 to 2014 in Taiwan. *J. Formos. Med. Assoc.* **118**, S66–S73 (2019).
3. Fletcher, B., Gulanick, M. & Lamendola, C. Risk factors for type 2 diabetes mellitus. *J. Cardiovasc. Nurs.* **16**, 17–23 (2002).
4. Chen, R., Ovbiagele, B. & Feng, W. Diabetes and stroke: Epidemiology, pathophysiology, pharmaceuticals and outcomes. *Am. J. Med. Sci.* **351**, 380–386 (2016).
5. Boehme, A. K., Esenwa, C. & Elkind, M. S. Stroke risk factors, genetics, and prevention. *Circ. Res.* **120**, 472–495 (2017).
6. Arboix, A. Cardiovascular risk factors for acute stroke: Risk profiles in the different subtypes of ischemic stroke. *World J. Clin. Cases* **3**, 418 (2015).
7. Wang, W. et al. A longitudinal study of hypertension risk factors and their relation to cardiovascular disease: The strong heart study. *Hypertension* **47**, 403–409 (2006).
8. Balakumar, P., Maung-U, K. & Jagadeesh, G. Prevalence and prevention of cardiovascular disease and diabetes mellitus. *Pharmacol. Res.* **113**, 600–609 (2016).
9. Mamdouh, H. et al. Prevalence and associated risk factors of hypertension and pre-hypertension among the adult population: Findings from the dubai household survey, 2019. *BMC Cardiovasc. Disord.* **22**, 18 (2022).
10. Sadr, H., Salari, A., Ashoobi, M. T. & Nazari, M. Cardiovascular disease diagnosis: A holistic approach using the integration of machine learning and deep learning models. *Eur. J. Med. Res.* **29**, 455 (2024).
11. Saberi, Z. A., Sadr, H. & Yamaghani, M. R. An intelligent diagnosis system for predicting coronary heart disease. In *2024 10th International Conference on Artificial Intelligence and Robotics*, 131–137 (2024).
12. Nazari, M., Emami, H., Rabiei, R., Hosseini, A. & Rahmatizadeh, S. Detection of cardiovascular diseases using data mining approaches: Application of an ensemble-based model. *Cogn. Comput.* **16**, 2264–2278 (2024).
13. Olorunfemi, B. O. et al. Efficient diagnosis of diabetes mellitus using an improved ensemble method. *Sci. Rep.* **15**, 3235 (2025).
14. García-Olmos, L. et al. Comorbidity patterns in patients with chronic diseases in general practice. *PLoS ONE* **7**, e32141 (2012).
15. Loza, E., Jover, J. A., Rodriguez, L. & Carmona, L. Multimorbidity: Prevalence, effect on quality of life and daily functioning, and variation of this effect when one condition is a rheumatic disease. *Semin. Arthritis Rheum.* **38**, 312–319 (2009).
16. Kim, G., Lim, H., Kim, Y., Kwon, O. & Choi, J.-H. Intra-person multi-task learning method for chronic-disease prediction. *Sci. Rep.* **13**, 1069 (2023).
17. Feng, R. *et al.* Chronet: A multi-task learning based approach for prediction of multiple chronic diseases. In *Multimedia Tools and Applications* 1–15 (2022).
18. Tsai, H. et al. Multimodal attention network for dementia prediction. *IEEE J. Biomed. Health Inf.* **28**, 6918–6930 (2024).
19. Rendle, S. Factorization machines. In *IEEE International Conference on Data Mining*, 995–1000 (2010).
20. Guo, H., Tang, R., Ye, Y., Li, Z. & He, X. Deepfm: A factorization-machine based neural network for ctr prediction. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1725–1731 (2017).
21. Lian, J. et al. xdeepfm: Combining explicit and implicit feature interactions for recommender systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1754–1763 (2018).
22. Wang, R., Fu, B., Fu, G. & Wang, M. Deep & cross network for ad click predictions. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1–7 (2017).
23. Wang, R. et al. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference*, 1785–1797 (2021).
24. Ruder, S. An overview of multi-task learning in deep neural networks. Preprint at https://arxiv.org/abs/1706.05098 (2017).
25. Baxter, J. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Mach. Learn.* **28**, 7–39 (1997).
26. Yang, Y. & Hospedales, T. Trace norm regularised deep multi-task learning. Preprint at https://arxiv.org/abs/1606.04038 (2017).
27. Tang, H., Liu, J., Zhao, M. & Gong, X. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, 269–278 (2020).
28. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at https://arxiv.org/abs/1301.3781 (2013).
29. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, vol. 26 (2013).
30. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
31. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
32. Birk, J. L. et al. Depression and multimorbidity: Considering temporal characteristics of the associations between depression and multiple chronic diseases. *Health Psychol.* **38**, 802 (2019).
33. Bobo, W. V. et al. Association of depression and anxiety with the accumulation of chronic conditions. *JAMA Netw. Open* **5**, e229817–e229817 (2022).

Author contributions

H.T., T-W.Y., and C-F.C. conceived the experiments. T-Y.W. and C-L.C. conducted the experiments. H.T. and T-W.Y. analyzed the results. Y-C.T. and C-F.C. contributed to data preparation. H.T., T-W.Y., and T-Y.W. wrote most of the manuscript. H.T., T-W.Y., and C-L.C. revised the manuscript. All authors reviewed the manuscript.

Funding

This research was supported in part by the National Science and Technology Council of Taiwan (NSTC 113-2221-E-002-201-), Ministry of Science and Technology of Taiwan (MOST 110-2622-8-002-018, MOST 110-2221-E-002-112-MY3), National Taiwan University (NTU 109-3111-8-002-002), and Cathay Life Insurance.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.T. or C.-F.C.