Informatics in Medicine Unlocked 30 (2022) 100928

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Sequential machine learning in prediction of common cancers

Jovan Andjelkovic [a], Branimir Ljubic [a,b], Ameen Abdel Hai [a], Marija Stanojevic [a], Martin Pavlovski [a], Wilson Diaz [a], Zoran Obradovic [a,*]

[a] Temple University, Center for Data Analytics and Biomedical Informatics (DABI), Philadelphia, PA, 19121, USA
[b] Rutgers University, The Office of Advanced Research Computing (OARC), Piscataway, NJ, 08854, USA

ARTICLE INFO

Keywords:
Lung cancer
Breast cancer
Cervix uteri cancer
Liver cell cancer
Machine learning

ABSTRACT

Cancer is one of the most common causes of death in the world. It is characterized by the multi-stage transformation of normal cells into tumor cells. Early cancer detection can significantly reduce its consequences, which was the objective of many machine learning (ML) published studies. However, most of them focused on microarray, gene expression, or publicly available medical datasets. Almost none offered an approach for predicting cancer through analysis of sequential data, such as Electronic Health Record (EHR) data.

This paper presents a sequential ML approach to predict the occurrence of lung cancer, breast cancer, cervical cancer, and liver cell cancer using EHR data. The accuracy of sequence learning models based on long short-term memory (LSTM) and bidirectional gated recurrent units (GRU) were compared to traditional ML methods based on multilayer perceptron, random forest, decision tree, and K-nearest neighbor. The models were trained and tested on 50,606 patient hospitalization histories. Unsupervised and supervised data reduction methods (singular value decomposition (SVD) and a neural network embedding layer) were applied to overcome the challenges of high-dimensionality and sparsity of EHR data.

The results provided evidence that for this application GRU outperforms alternatives based on accuracy, Area Under the Receiver Operating Characteristic curve (AUROC), sensitivity (recall), specificity, precision, and F1 score. It was the best performing model with accuracy between 81% (breast cancer) and 88% (liver cancer) on balanced out of sample EHRs. Multilayer perceptron and LSTM manifested comparable performances (accuracies between 78% and 87%) among the alternatives, while decision tree was the worst-performing model.

The findings of this study could potentially aid medical professionals in cancer diagnostics, treatment, and prevention. In particular, experiments confirmed that GRU could accurately predict cancer by learning from simplified patient representations using an embedding layer or SVD. Therefore, GRU's predictions could be used in early cancer detection, potentially improving patients' survival rates.

1. Introduction

Cancer is one of the deadliest diseases of the 21st century. It was proclaimed a leading cause of death worldwide in 2020 (nearly 10 million deaths), the same year when the COVID-19 pandemic started (5.53 million deaths) [1]. One of the reasons this disease is so devastating is that it affects individuals of all ages, races, and genders. Additionally, differences in genetics, environmental and other factors can lead to differences in developing cancers among different people. More than 100 cancer types have been discovered, mostly named for the organs or tissues where cancer originates. Breast cancer is the most common cancer among women, with approximately 2 million new cases over the globe every year [2]. About 2 million new lung cancer cases and more than 800,000 liver cancer patients appear worldwide every year [2,3]. Around 550,000 new cases of cervical cancer are reported worldwide each year, and it occurs most often in women over age 30 [2].

Machine learning (ML) models have been used for the prediction and detection of different medical conditions [4,5], including various cancer types [6,7]. Many of them used genetic, molecular, and imaging data to predict the occurrence of cancer and optimal therapeutic approaches [8,9]. A support vector machine (SVM) was developed to predict the responses of 175 cancer patients to a variety of chemotherapeutic drugs [10]. The authors evaluated the models on the gene-expression profiles of individual patient tumors, and the accuracies ranged from 81.5% to 82.6%. Microarray gene expression was examined for lung cancer [11]. The author preprocessed the data and trained multilayer perceptron to

[*] Corresponding author. Temple University. 1925 N 12thStreet, SERC 386, 035-10, Philadelphia, PA, 19122, USA.
E-mail address: zoran.obradovic@temple.edu (Z. Obradovic).

https://doi.org/10.1016/j.imu.2022.100928
Received 29 November 2021; Received in revised form 5 March 2022; Accepted 21 March 2022
Available online 24 March 2022

determine the subset of genes most likely to cause cancer. Medical image analysis is also essential in today's medicine, but it requires a significant set of labeled data and appropriate ML techniques to make accurate predictions. A convolutional neural network (CNN) was constructed to classify hepatic tumor entities on multiphasic MRI [12]. The goal was to identify the correct radiological features present in test lesions. The dominating power of CNNs in medical fields was also confirmed by predicting lung cancer from CT imaging data [13]. The CNN architecture used in that study achieved a classification performance of around 0.9 AUC. With the increase of available medical data, the application of ML models expanded. A liver cancer study on 2890 patients used Cox multivariate regression for analyzing independent risk factors and artificial neural networks (ANNs) for prediction [14]. The ANNs achieved AUROC between 0.86 and 0.88 in 1-year survival analysis task. Another liver cancer study utilized the power of SVM to distinguish liver cancer patients from others [15]. Experiments aimed at predicting breast cancer using several traditional ML models showed that SVM was the best performing model with an accuracy of 0.979 [16]. However, despite the expansion of ML cancer papers in previous years, hardly any study focused on exploiting heterogeneous EHRs, although such data is updated with many new patient information daily and exists in almost every hospital.

This study proposes an ML approach to exploiting EHR data for cancer prediction. We predicted the occurrence of four common cancers: breast, cervical, liver, and lung cancer, using data from the Healthcare Cost and Utilization Project State inpatient database (HCUP SID) of California, collected between 2003 and 2011 [17]. The goal was to predict whether the patient would develop cancer within nine years of the first recorded hospitalization, learning from ICD-9 diagnosis codes. We hypothesized that two RNN models based on LSTM [18] and GRUs [19] achieve better prediction accuracy of cancer occurrence, as compared to four traditional ML models: multilayer perceptron (MLP) [20], random forest (RF) [21], decision tree (DT) [22] and K-nearest neighbor (KNN) [23]. Moreover, we hypothesized that expanding the sequence of hospitalizations at some point does not lead to accuracy improvement due to the curse of dimensionality and data sparsity. To test the hypothesis, we predicted cancers using different lengths of sequences by considering: up to 5, up to 10, up to 25, and up to 50 hospitalizations. In addition, we compared the efficacy of two different embedding methods (SVD and neural network embedding layer) on the lung cancer dataset, our most extensive dataset. Contributions of this study are the following:

- We successfully predicted four different cancers using EHR data alone, which is far more complicated to achieve than when imaging and/or gene data is available. Finding an appropriate approach for EHR representation (e.g. data extraction and preprocessing) is complex because such datasets are sequential, heterogeneous, and not stored for research purposes.
- Data reduction methods proved to be very useful in predicting cancers from highly dimensional and sparse data. They allowed us to summarize thousands of different diagnoses effectively from patients' histories without discarding the importance of any of them, instead of applying complex feature selections used in many previous studies.
- We showed that learning from longer sequences of hospitalizations does not necessarily imply higher prediction accuracy. In particular, when many patients have short hospital histories, considering more hospitalizations causes more padding with zeros. Adding a large fraction of zeros makes the data irrelevant, meaning that data reduction methods cannot help enough in summarizing valuable insights.

The rest of the paper is organized as follows: The Related work section provides a literature survey of recently published papers. The Materials and Methods section contains a detailed explanation of our EHR-based approach for predicting cancers (e.g. data preparation, dimensionality reduction methods). The obtained findings are presented in the Results section and analyzed in the Discussion section. Final insights were discussed in the Conclusion.

2. Related work

Risks of cancer development were extensively analyzed and large variability is reported between studies. Weegar and Sundstrom [24] predicted cervical cancer using free-text notes, diagnosis codes, drug codes, procedures, and lab results extracted from Swedish EHR data. It turned out that the clinical entities they retrieved from free text records provided the most precise predictions. However, such an approach leads to biased prediction models because phrases like "suspected cancer" usually appear in the notes before the diagnosis code is assigned. Atrey et al. [25] tried a dominance-based filtering approach to find the most important features for predicting breast cancer. They achieved a high accuracy between 98.9% and 99.6% by applying an ANN with only a few dominant features from the Wisconsin Breast Cancer dataset (WBCD). Although this filtering approach may be helpful, a drawback is that the approach was evaluated on only 699 patients and is missing a larger-scale prospective evaluation. Similarly, Li and Chen [26] used WBCD and another small dataset to apply several traditional ML models such as DT, RF, and SVM, without data preprocessing (e.g. feature selection, feature normalization). Moreover, both papers [25,26] based their predictions on only ten cytological attributes. Another breast cancer study introduced an ML-based decision support system, combined with random optimization for classifying primary breast cancer patients into two risk groups of progression [27]. The authors developed and applied the model to a sample of 454 patients. Besides achieving lower performances on a significantly smaller dataset, another drawback of this paper is focusing on only one institution. On the other hand, we created the datasets using the HCUP SID database that included more than 95% of hospital discharges in the US and inpatient care records from multiple participating states [17].

Two papers used publicly available datasets of 165 and 535 clinical patients for hepatocellular carcinoma (HCC) survival analysis, the most common kind of liver cancer [28,29]. It was a binary classification problem in both cases: whether the patient will die (0) or survive (1), indicated by a one-year outcome evaluation [28] or after the surgery [29]. RF achieved the highest prediction accuracy in both studies (80.64% and 73.9%).

Another liver disease study showed that the J48 decision tree algorithm could be beneficial in this topic with 0.507 mean absolute error [30]. The authors used the Indian Liver Patient Dataset, which contains 583 patients. Like in most previously mentioned papers, the predictions were also based on a small number of attributes. Yuan et al. [31] did a thorough lung cancer analysis (classification and survival analysis) using a dataset of 76,643 patients. The authors used multiple NLP techniques to extract essential features from structured and unstructured data. Even though the whole data preprocessing procedure was complex, and the dataset was significantly larger than ours, the final lung cancer classification model reached an AUROC of 92.7%. In contrast, our approach achieved around 92.2% without relying on NLP, which could lead to an optimistic estimate of the error due to possible information leaks from doctor and nurse notes. Miotto et al. used a complex autoencoder architecture to capture a compact and general-purpose set of features from the EHR data [32]. Instead of applying deep learning on the preprocessed EHR data, they used deep sequence learning to generate better representations of more than 700, 000 patients to predict liver cancer along with 77 other diseases. Another deep learning ensemble method was used for feature extraction from EHRs of 1000 patients to predict lung cancer treatment failure [33]. The problem was formulated as binary classification, where the authors predicted lung cancer readmission using an ensemble of MLP models and model selection with adaptive multi-objective optimization.

Choi et al. [34] proposed an approach for learning the hierarchy between diagnosis and procedure codes. They experimented with three different deep learning models to learn the multilevel embedding representation of a patient. Additionally, in Ref. [35], the authors applied similar learning techniques to monitor ICU mortality risk. By combining the Latent Semantic Analysis (LSA) and LSTM, they learned a representation of the embedding sequence using laboratory tests, vital signs, and medications. Both papers [34,35] used RNN in combination with logistic regression, where the former was utilized only for representation learning, while the latter was used for binary classification. However, those two mentioned approaches are different compared to our objective. We used SVD and an embedding layer for representation learning, while RNNs were applied for cancer prediction.

Most of the mentioned studies focused on genetic, imaging, or publicly available medical datasets, and almost none predicted cancer from sequential EHR data. Additionally, almost all prior studies were evaluated on much fewer examples than we used (except [31,32]).

3. Materials and methods

The proposed ML methods were trained on EHR data extracted from the HCUP SID California database. This database has the most extensive collection of longitudinal hospital care data in the US, with encounter-level data dating back to 1988 [17]. SID is one of the HCUP databases created in 1995, encompassing inpatient discharge abstracts from participating states. It contains clinical and nonclinical variables such as principal and secondary diagnoses, procedures, admission and discharge status, patient demographics characteristics, etc. We created a dataset of nine years (2003–2011) from this database, containing all hospitalizations of patients in participating hospitals.

We extracted and preprocessed the data for each cancer separately, focusing on diagnoses only. The format of all diagnoses codes follows the "International Classification of Diseases, Ninth Revision" (ICD-9). After this step, we created four datasets, one for each cancer. The main challenges in preprocessing were a heterogeneous number of hospitalizations across patients and a high dimensionality of possible ICD-9 diagnoses in each hospitalization. We observed all hospitalizations before a particular cancer was first diagnosed for positive patients (diagnosed with particular cancer). We considered only patients who had a minimum of four hospitalizations before the diagnosis of cancer because it has been shown previously that RNN models perform optimally with sequences of four or more hospitalizations [4]. We randomly chose an equal number of negative (cancer-free) patients from the original HCUP database. The group of negative patients also had at least four hospitalizations. We balanced the average length of sequences to ensure the similarity between positive and negative classes. The number of patients and the number of ICD-9 diagnoses included in the final datasets are presented in Table 1.

Furthermore, we balanced all four datasets to ensure that each has similar mean age in positive and negative cohorts (Table 2).

The preprocessed dataset for each cancer can be represented as a sparse matrix, where the rows represent patients. Each row contains a patient's hospitalizations in the sequential order in which the patient's visits occurred. Different patients had a different number of hospitalizations before the cancer was diagnosed. Therefore, we observed up to 50 visits for each patient. If the patient had less than 50 hospitalizations, we filled the rest with zeros. We used a one-hot encoded (also known as a "bag-of-words") representation. Each value was set to 1 or 0, indicating whether a disease was diagnosed at a specific visit or not, respectively. Most of the values were zeros since only a few diagnoses were found during each hospitalization.

Table 1
The total number of patients and diagnoses ICD9 codes in four cancer datasets.

Cancer type	Total # of patients	Number of distinct disease ICD9 codes
Lung cancer	28,038	7024
Breast cancer	12,498	6244
Cervix uteri	1748	4166
Liver cell cancer	8322	5713

Table 2
Patients' statistics for four cancer datasets.

Cancer type	Patient class	Average age
Lung cancer	Positive	73
	Negative	75
Breast cancer	Positive	71
	Negative	71
Cervix uteri	Positive	59
	Negative	60
Liver cell cancer	Positive	65
	Negative	67

Given that there were many diagnoses in the one-hot encoded representation (as evident in Table 1), we combined SVD with each prediction model to reduce data dimensionality. As an alternative, we also applied embedding layers in RNN models for data dimensionality reduction. Both approaches are described in the following subsection.

3.1. Dimensionality reduction

To choose relevant features and reduce the dimensionality (i.e. the number of features), we experimented with two embedding methods:

- Singular value decomposition (SVD)
- Incorporating an embedding layer in the architecture of RNN models

Singular value decomposition is a well-developed method for extracting dominant features of large datasets and reducing data dimensionality [36]. It generates *unsupervised* embeddings since it is unaware of the prediction target during the generation process. Given that traditional ML models cannot capture the sequential correlation between hospitalizations like RNN models, we had to create different inputs for RNN and traditional ML models. We applied SVD separately on each of the inputs (Fig. 1).

For RNN models, each cancer dataset was represented as a 3D matrix (tensor) C of size $p \times h \times d$, where p is the number of patients, h is the number of hospitalizations, and d is the number of unique diagnoses. For example, preprocessed lung cancer dataset had the following dimensions: 28,038 × 50 × 7024. For each of the 28,038 patients, we recorded the first 50 hospitalizations, where for each hospitalization, there were 7024 different diagnoses. After applying SVD, the result was a lower rank tensor C of size $p \times h \times r$, where r was much smaller than d. The optimal number of reduced components r was determined based on Experiment 1 shown in the Results section. We experimented with different values between 100 and 500. When we chose 500, the dimensions of the lung cancer dataset were reduced to the following: 28,038 × 50 × 500. On the other hand, for traditional models, we preprocessed 2D matrices T of size $p \times d$. In this format, the lung cancer dataset had the following dimensions: 28,038 × 7024. Each value in such matrix represented how many times a particular diagnosis was found across the first 50 patient hospitalizations. Additionally, we scaled all the values to the range between 0 and 1 before using SVD. The SVD output was a matrix T of size $p \times r$. When we set $r = 500$, the lung cancer dataset for traditional models was reduced to 28,038 × 500. In both cases, RNN and standard inputs, reduced dimensionality matrices still carry the essential information from their respective original representation.

Another compelling feature selection method based on an embedding layer was incorporated in the RNN model. Embedding layers are designed for learning vector representations of categorical data [37]. In contrast to SVD, an RNN embedding layer produces *supervised* embeddings as a part of the training process. Additionally, this layer is trained

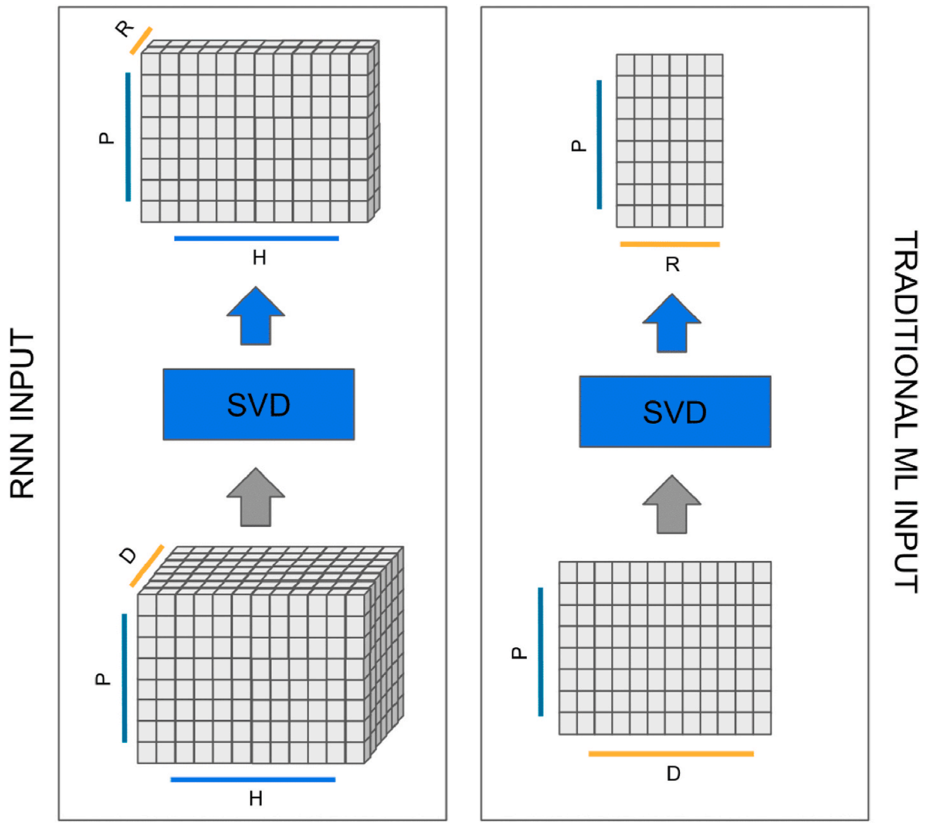

Fig. 1. An application of the SVD reduction method on RNN input (left) and standard ML input (right).

like all other layers in the network architecture, i.e. to minimize the loss function using the selected optimization method.

We used an embedding layer for the same purpose as SVD, which was to map the categorical input variables to real-valued vectors of smaller dimensionality than the input vectors. For example, we utilized an embedding layer to map a one-hot representation of the mentioned lung cancer dataset (28,038 × 50 × 7024) to a real-valued 28,038 × 50 × 500 matrix. In other words, an embedding layer learned (during the training process) a 50 × 500 embedding representation for each of the 28,038 patients. The learned embeddings were then used as feature vectors for predictions.

3.2. Prediction models

The prediction problem is formulated as binary classification. The hospitalization when cancer occurred was used as a class label. If diagnosed with cancer, we assigned a patient to the positive class (‘1′). Otherwise, we put the patient into the negative class (‘0′). We experimented with two different RNN models. These models are advantageous for the sequence data, especially when one data point is dependent on the preceding data point, like in our case. The reason is that they have a memory to store the states or information of previous inputs in order to construct the sequence’s subsequent output. This mechanism is also known as a hidden state. The following equations explain the learning process:

$$h_{t+1} = relu(W_x X_t + W_h X_t + b_h) \tag{1}$$

$$y_{t+1} = sigmoid\left(W_y * h_{t+1} + b_y\right) \tag{2}$$

To calculate the hidden state h_{t+1} for the next step $t+1$, we use input weights W_x and hidden units weights W_h together with the input X_t from the current time step t, and bias b_h from the recurrent layer. At the end of the calculation, a nonlinear transformation ReLU is applied. Furthermore, to predict y_{t+1}, we multiply the newly learned hidden state with the weights W_y from the output layer. We also add up bias b_y of all neurons in the network. Finally, everything is pulled through a sigmoid function.

The first model contains layers with LSTM units capable of learning long-term dependencies in sequential data. Remembering information for long periods is practically their default behavior. The second model has layers with GRUs. Unlike the LSTM unit, the GRU has gating units that modulate information flow without separating memory cells [38]. This structure allows to adaptively capture dependencies from large data sequences without discarding information from earlier parts of the sequence.

The architectures of both models are identical, with one hidden layer of 64 neurons (Fig. 2). Empirical evaluation of RNN models showed that both the LSTM and GRU demonstrated superiority over traditional ML models [39]. Since LSTM and GRU architectures have shown surpassing results in various applications, we compared both in our experiments.

SVD and embedding layer were tested separately with both RNN methods. The output layer contains only one neuron with the sigmoid activation function. The adaptive learning rate optimization algorithm ADAM was used to train the RNN models [40].

A potential problem with training neural networks could be the number of epochs. A large number of epochs could lead to overfitting, whereas an insufficient number of epochs may result in an underfit model. That is why in our application, sequential learning models used the early stopping method, which monitored the model’s performance during training. The objective of the method is to stop the training when the validation loss (binary cross-entropy loss) starts to increase constantly. As a result, both RNN models were trained through 20 epochs unless stopped earlier by the method mentioned above.

We used a batch size of 64 since, in such a way, the overall training procedure required less memory. Furthermore, a smaller size was chosen because it is reported across many applications that using such small

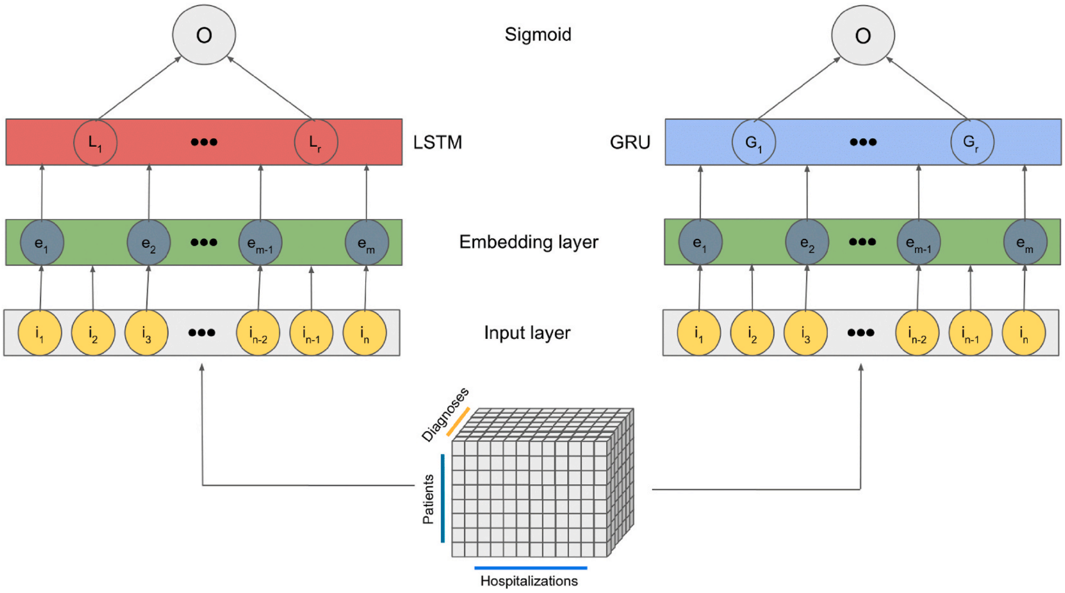

Fig. 2. Architectures of RNN models: LSTM (left) and Bidirectional GRU (right).

batch sizes achieves training stability and improved generalization performance [41].

To compare the performance of the proposed sequence learning models, we also trained four standard machine learning models: DT, MLP, RF, and KNN. Only default settings provided in the scikit-learn Python library were used for DT and MLP without parameter tuning [42]. For RF and KNN, we used the standard implementation with basic settings (for RF the maximum depth was set to 10 and the number of trees to 100, while for KNN the number of nearest neighbors was 3). All prediction models were run separately for each of the four studied cancers. We trained the models on 80% of patients selected entirely at random, and the remaining 20% we used for testing, while 25% of the training set was used for training validation. All models were run on balanced datasets, and we measured test accuracy, Area Under the Receiver Operating Characteristic curve (AUROC), sensitivity (recall), specificity, precision, and F1 score. Prediction accuracy was chosen as a primary metric since there are equal patients in both classes for each cancer. However, we also reported the AUROC score for a more comprehensive evaluation of the models. The difference between these two metrics is based on the decision threshold, i.e. class probability threshold. In binary classification, the threshold is the value over which a sample is assigned to class one. AUROC is a metric that evaluates a binary classifier's output over decision thresholds varying between 0 and 1, whereas the accuracy indicates how well a classifier performs for the default threshold of 0.5. High accuracy and high AUROC indicate that the classifier performs admirably for the default threshold and similarly for many other threshold values. Additionally, an admirably accurate classifier should have high sensitivity and specificity. Since the AUROC score summarizes the model's efficacy in terms of sensitivity and specificity for various decision thresholds, we calculated those two metrics only for the 0.5 threshold. The source code is available at the following Github repository: https://github.com/jovanandj/ML_cancers_prediction.

4. Results

4.1. Comparison of embedding methods

We used our most extensive dataset, the lung cancer dataset, to compare the embedding methods. A comparison of the predictive

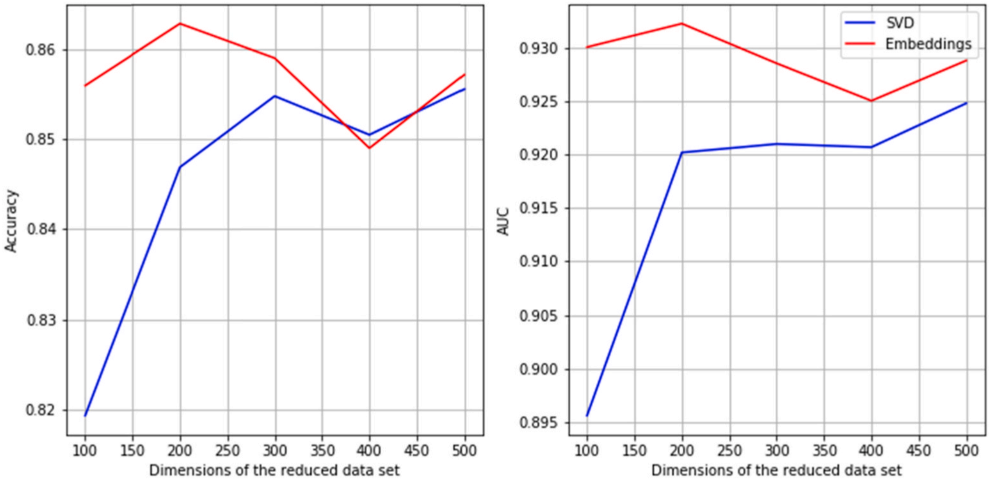

Fig. 3. GRU's accuracy (left) and AUROC (right) with: SVD (blue) or RNN Embedding layer (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

performances obtained when using SVD or an embedding layer is shown in Fig. 3.

In general, the two methods acted very similarly across different dimensions. The most significant difference in performance can be seen when the data was reduced to 100 dimensions, where an Embedding layer showed a few percent higher performances than SVD. However, as the number of dimensions kept increasing, SVD produced inputs that led to comparable performances as Embedding layer' inputs. In fact, at 400 dimensions, SVD was a little bit better in helping GRU predict lung cancer based on accuracy.

4.2. Experiments with cancer datasets

The results of five experimental runs in terms of accuracy, AUROC score, sensitivity (recall), specificity, precision, and F1 score are reported in Table 3. We labeled the best results in bold for each type of cancer.

Comparisons of all six prediction models for each cancer dataset are shown in Figs. 4 and 5. Given that SVD led to better prediction performance when it reduced the data dimensionality to 300 and 500 features, we chose the lowest dimension that performed the best. We trained all models on datasets with 300 SVD-preprocessed features. Additionally, all datasets were balanced, containing 50% of cancer-positive patients.

GRU generally obtained the best performance across the six evaluation metrics on all datasets, with 2–18% higher accuracy and a considerably higher AUROC score (between 0.87 and 0.94) than all alternatives. This model showed the highest accuracy of 88% in predicting liver cancer. Additionally, it can be observed that GRU made excellent predictions for the 0.5 decision threshold, while for other thresholds, it may perform even better. Interestingly, LSTM and MLP achieved comparable accuracy in predicting cancers (between 78% and 87%), even though LSTM was given more informative input. Moreover, MLP considerably outperformed LSTM in terms of AUROC. As for the other traditional ML models, DT was the worst-performing classifier on all four datasets with an accuracy between 63% and 71%.

In summary, all six models achieved the best results overall when predicting liver cancer, while the most challenging classification task was breast cancer prediction. Also, all models have shown admirable performances in predicting lung cancer.

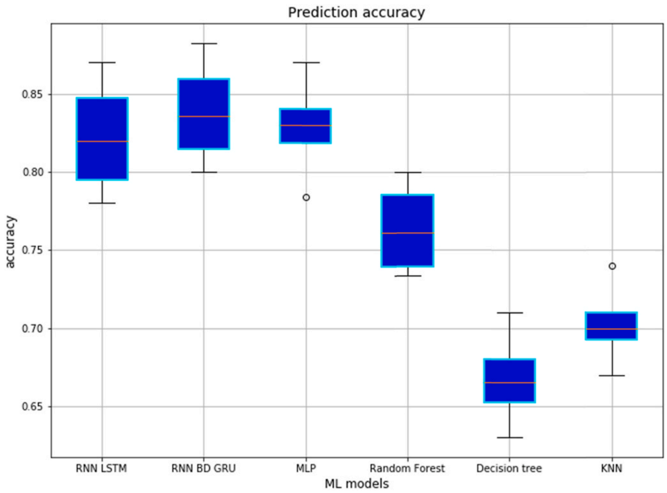

Fig. 4. Accuracy of six ML models on four types of cancer.

4.3. Experimenting with different lengths of sequences

Besides the experiments based on all available visits in our datasets, we also wanted to test models' performance by observing fewer visits. In particular, our goal was to infer a sufficient number of observed visits for reaching optimal predictions, and we wanted to estimate the usefulness of padding. The number of patients was constant through this experiment, focusing on different numbers of earliest visits: up to 5, up to 10, up to 25, and up to 50. The choice of these four sequences' lengths was based on the statistics shown in Table 4. Since the average number of visits was between seven and nine, we tested our models on 5-visits and 10-visits sequences because we expected the biggest padding increase between these two lengths. Moreover, we had to pick another sequence length closer to 50 visits, representing most of the patients from the datasets. That is why we also tried with 25-long sequences.

Different lengths of visit sequences require more or less padding (filling in missing data with zeros, as described in the Methodology section), depending on how many visits each patient in the dataset had. For example, there was far less padding in "up to 5″ than in the "up to

Table 3
Results of sequential learning (LSTM and GRU) and traditional models (MLP, RF, DT, KNN) trained on the four cancer datasets after creating 300 SVD features for each dataset. The results are reported in terms of the models' mean performances over five runs and the corresponding standard deviations.

Type of cancer	Models	Accuracy	AUROC	Sensitivity (Recall)	Specificity	Precision	F1 score
Lung cancer (D162)	GRU	**0.85 ± 0.005**	**0.92 ± 0.001**	**0.86 ± 0.02**	**0.84 ± 0.02**	**0.84 ± 0.01**	**0.85 ± 0.004**
	LSTM	0.84 ± 0.004	0.75 ± 0.01	0.74 ± 0.01	0.62 ± 0.02	0.66 ± 0.01	0.7 ± 0.004
	MLP	0.83 ± 0.004	0.83 ± 0.004	0.83 ± 0.01	0.83 ± 0.01	0.83 ± 0.005	0.83 ± 0.01
	RF	0.78 ± 0.01	0.78 ± 0.01	0.77 ± 0.005	0.79 ± 0.01	0.79 ± 0.01	0.78 ± 0.004
	DT	0.67 ± 0.01	0.67 ± 0.01	0.67 ± 0.01	0.67 ± 0.01	0.67 ± 0.01	0.67 ± 0.01
	KNN	0.7	0.7	0.83 ± 0.004	0.58 ± 0.004	0.66 ± 0.005	0.74 ± 0.005
Breast cancer (D174)	GRU	**0.81 ± 0.01**	**0.87 ± 0.004**	**0.81 ± 0.023**	0.77 ± 0.02	**0.79 ± 0.01**	0.8 ± 0.01
	LSTM	0.78 ± 0.01	0.7 ± 0.01	0.73 ± 0.04	0.53 ± 0.05	0.61 ± 0.02	0.67 ± 0.01
	MLP	0.78 ± 0.01	0.78 ± 0.01	0.79 ± 0.02	**0.78 ± 0.01**	0.78 ± 0.01	0.79 ± 0.01
	RF	0.74 ± 0.01	0.74 ± 0.01	0.7 ± 0.01	**0.78 ± 0.01**	0.76 ± 0.01	0.73 ± 0.01
	DT	0.63 ± 0.01	0.63 ± 0.01	0.64 ± 0.01	0.62 ± 0.02	0.63 ± 0.02	0.64 ± 0.01
	KNN	0.67 ± 0.004	0.67 ± 0.004	0.8 ± 0.01	0.53 ± 0.004	0.63 ± 0.01	0.71 ± 0.01
Cervix uteri cancer (D180)	GRU	0.82 ± 0.01	**0.89 ± 0.03**	0.79 ± 0.07	**0.82 ± 0.03**	**0.81 ± 0.02**	0.8 ± 0.04
	LSTM	0.8 ± 0.01	0.7 ± 0.02	0.73 ± 0.07	0.53 ± 0.03	0.6 ± 0.02	0.66 ± 0.03
	MLP	**0.83 ± 0.01**	0.83 ± 0.01	**0.84 ± 0.02**	0.81 ± 0.02	**0.81 ± 0.02**	**0.82** ± 0.01
	RF	0.73 ± 0.02	0.73 ± 0.02	0.68 ± 0.03	0.79 ± 0.01	0.75 ± 0.02	0.71 ± 0.02
	DT	0.66 ± 0.02	0.66 ± 0.02	0.68 ± 0.03	0.63 ± 0.05	0.64 ± 0.02	0.66 ± 0.01
	KNN	0.7 ± 0.01	0.71 ± 0.01	0.82 ± 0.02	0.59 ± 0.03	0.66 ± 0.01	0.73 ± 0.01
Liver cancer (D155.0 and D155.1)	GRU	**0.88 ± 0.005**	**0.94 ± 0.004**	**0.87 ± 0.01**	**0.87 ± 0.02**	**0.87 ± 0.01**	**0.87 ± 0.01**
	LSTM	0.87 ± 0.004	0.82 ± 0.006	0.75 ± 0.01	0.72 ± 0.01	0.73 ± 0.01	0.74 ± 0.004
	MLP	0.87 ± 0.01	0.87 ± 0.01	**0.87 ± 0.01**	**0.87 ± 0.005**	**0.87 ± 0.01**	**0.87 ± 0.004**
	RF	0.8 ± 0.005	0.8 ± 0.004	0.74 ± 0.01	0.86 ± 0.01	0.84 ± 0.01	0.79
	DT	0.71 ± 0.01	0.71 ± 0.01	0.71 ± 0.01	0.69 ± 0.03	0.7 ± 0.01	0.71 ± 0.01
	KNN	0.74 ± 0.01	0.74 ± 0.01	0.81 ± 0.01	0.68 ± 0.01	0.72 ± 0.01	0.76 ± 0.01

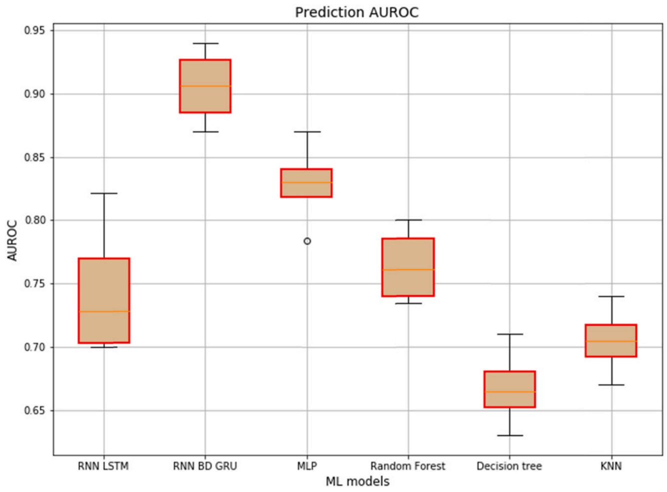

Fig. 5. AUROC score of six ML models.

Table 4
Statistics of visits/hospitalizations separated into four groups: up to 5, up to 10, up to 25, and up to 50, across four cancer datasets. The average number of visits was reported regarding the mean number of visits and the corresponding standard deviation.

Cancer type	Average number of visits	Number of visits	Number of patients	Padding percent
Lung cancer	7.4 ± 4.9	Up to 5	12,824	5.4
		Up to 10	23,638	35.5
		Up to 25	27,739	71.0
		Up to 50	28,038	85.3
Breast cancer	7.5 ± 5.0	Up to 5	5713	5.6
		Up to 10	10,450	35.4
		Up to 25	12,340	70.5
		Up to 50	12,498	85.0
Cervix uteri cancer	9.1 ± 7.8	Up to 5	684	4.5
		Up to 10	1326	31.1
		Up to 25	1683	65.7
		Up to 50	1748	82.1
Liver cancer	7.81 ± 5.3	Up to 5	3560	5.3
		Up to 10	6765	33.8
		Up to 25	8192	69.4
		Up to 50	8322	84.4

50″ experiment for the same 28,038 patients of the lung cancer dataset. The reason was that patients with long hospital histories were rare. Therefore, in the “up to 5″ lung cancer experiment, the models analyzed all 28,038 patients, which included 12,824 patients with up to five visits and 15,214 with longer hospital histories but using only the first five visits for the patients with more than five visits. This means we applied padding only for those who had less than five visits. We hypothesized that increasing the visits would not constantly improve the model's accuracy.

The results showed that most models had similar performances on 25 and 50 visits, implying that the models saturated when they observed more than 25 visits. For example, GRU learned well and performed identically with both sequence sizes in all four prediction tasks. Its accuracy was between 81% (breast cancer) and 88% (liver cancer). In general, the models' accuracy difference between these two sequence lengths was 1% on average. Moreover, the difference between 10 and 25 hospitalizations was more significant, around 1–4%, suggesting that models could achieve decent performances using ten visits, but not as optimal as learning from 25 visits.

On the other hand, all models produced the worst predictions when they observed only five visits. On average, the accuracy was 4% less than the accuracy achieved with ten visits. Additionally, this accuracy difference between 5-visit and 10-visit sequences turned out to be the most considerable accuracy increase. The worst performing models in this experiment were DT and KNN. DT accuracy was between 59% and 73%, while for KNN, it was in the range of 61%–74%.

5. Discussion

This study compared sequential learning RNN and traditional ML models using diagnosis codes extracted from EHR data for four common cancers. The obtained results (Fig. 3) showed that the embedding layer and SVD were comparable in finding relevant features in 7024 different diagnoses from the lung cancer dataset. In some cases, the GRU model achieved slightly better accuracy and AUROC score using an embedding layer. A possible explanation for the outcome is that the RNN embedding layer takes an integral part of the recurrent network architectures. Thus, backpropagation allows for learning embeddings tailored to the cancer prediction task. On the other hand, SVD is a separate (out-of-architecture) embedding method that produces low-dimensional embeddings before the model training. In that regard, the SVD-based embeddings are learned in an unsupervised manner, making them task-agnostic, and thus, they are not specialized for cancer prediction.

The second experiment (Table 3) compared all six ML models on every cancer dataset. GRU outperformed the other methods on all datasets, except the cervical cancer dataset. For Cervix uteri, MLP showed better performance than both RNN models. The reason might be the smallest dataset size of only 1748 patients. Thus, insufficient training samples made it difficult for sequential learning models to find the temporal correlation between patient diagnoses. Another critical point might be that the number of unique diagnoses in the original cervical cancer dataset was almost five times larger than the number of patients, meaning significantly more features than examples. Although we applied SVD, which is considered suitable for dimensionality reduction under such conditions, the insufficient number of examples may still be an obstacle in sequential learning.

All six models achieved the best results on the liver cancer dataset. Additionally, lung cancer prediction results were also admirable. Although we did not use age in preprocessing to balance the cohorts' distribution, the patients' age statistics might adequately explain this outcome. In particular, the most significant age difference between negative and positive patients was in lung and liver cell datasets, meaning that the distribution of the cohorts was a bit different in these datasets. Thus, it might be simpler for classifiers to learn to distinguish patients.

The arguments for this assumption could be found by analyzing specificity. The highest specificity (which measures the proportion of correctly identified negative patients) for all models was achieved primarily in lung cancer and liver cancer predictions. Considering that negative patients in these two datasets were two years older on average, this might be why the models had better performances in those two prediction tasks. Moreover, this observation may also be supported by the fact that the worst performance was obtained when predicting breast cancer, for which the mean age in both cohorts was identical. However, we need to emphasize that the patient's age was not explicitly utilized as a prediction feature.

In general, GRU was better than the other models with respect to six evaluation metrics. Judging by the prediction accuracy, GRU performed at least 2% better in every case, while in terms of AUROC, it was significantly better than alternatives. On the other hand, the worst model was DT, which was expected considering its modeling capacity. It allows for neither learning abstract patient representations nor capturing any temporal correlations between patients' diagnoses. The LSTM model performed similarly to MLP, even though LSTM was given additional temporal information.

In the third experiment (Table 5), we estimated models' accuracy by observing up to 5, up to 10, up to 25, and 50 hospitalizations. On 25 and 50 visits, most of the models behaved similarly, while on up to 5 visits, all models achieved the lowest accuracies. The largest significant gain in

Table 5
Accuracy of the models trained with 5, 10, 25, and 50 visits/hospitalizations sequences. The results are reported regarding the models' mean performances over five runs and the corresponding standard deviations.

Cancers	visits	GRU	LSTM	MLP	RF	DT	KNN
Lung cancer	5	**0.8 ± 0.003**	0.75 ± 0.002	0.77 ± 0.003	0.75 ± 0.005	0.63 ± 0.01	0.61 ± 0.01
	10	**0.84 ± 0.003**	0.79 ± 0.003	0.82 ± 0.005	0.77 ± 0.004	0.66 ± 0.005	0.67 ± 0.01
	25	**0.85 ± 0.002**	0.83 ± 0.005	0.83 ± 0.002	0.78 ± 0.01	0.68 ± 0.005	0.63 ± 0.005
	50	**0.85 ± 0.01**	0.84 ± 0.004	0.83 ± 0.003	0.78 ± 0.005	0.67 ± 0.01	0.7
Breast cancer	5	**0.75 ± 0.003**	0.71 ± 0.004	0.74 ± 0.01	0.72 ± 0.003	0.61 ± 0.01	0.61 ± 0.01
	10	**0.79 ± 0.01**	0.74 ± 0.01	0.77 ± 0.004	0.74 ± 0.004	0.64 ± 0.01	0.63 ± 0.01
	25	**0.81 ± 0.004**	0.78 ± 0.004	0.79 ± 0.002	0.74 ± 0.01	0.64 ± 0.01	0.63 ± 0.005
	50	**0.81 ± 0.007**	0.78 ± 0.005	0.78 ± 0.01	0.74 ± 0.01	0.64 ± 0.01	0.67 ± 0.01
Cervix uteri cancer	5	**0.76 ± 0.02**	0.71 ± 0.02	**0.76 ± 0.02**	0.69 ± 0.02	0.59 ± 0.02	0.62 ± 0.02
	10	**0.81 ± 0.01**	0.76 ± 0.01	0.79 ± 0.02	0.73 ± 0.02	0.62 ± 0.03	0.64 ± 0.01
	25	**0.82 ± 0.02**	0.79 ± 0.02	0.78 ± 0.03	0.71 ± 0.02	0.64 ± 0.01	0.62 ± 0.02
	50	0.82 ± 0.01	0.8 ± 0.01	**0.83 ± 0.01**	0.73 ± 0.02	0.66 ± 0.02	0.7 ± 0.01
Liver cancer	5	**0.84 ± 0.01**	0.8 ± 0.01	**0.84 ± 0.01**	0.78 ± 0.01	0.68 ± 0.01	0.7 ± 0.01
	10	**0.87 ± 0.004**	0.83 ± 0.005	0.86 ± 0.01	0.81 ± 0.005	0.7 ± 0.01	0.73 ± 0.01
	25	**0.88 ± 0.004**	0.86 ± 0.004	0.87 ± 0.01	0.82 ± 0.01	0.73 ± 0.01	0.73 ± 0.01
	50	**0.88 ± 0.005**	0.87 ± 0.004	0.87 ± 0.01	0.8 ± 0.005	0.71 ± 0.01	0.74 ± 0.01

accuracy was reported between 5-visit and 10-visit sequences. Furthermore, some models showed even better results with shorter input sequences, such as DT and liver cancer dataset or MLP and breast cancer dataset. These outcomes are probably related to the size of the dataset, the model's learning power, and the complexity of the prediction task. For instance, all six models predicted liver cancer the same as with 50 visits or better when the input sequences contained up to 25 visits, potentially implying that this task was undemanding with an acceptable amount of information in the first 25 visits.

This paper showed a successful approach for utilizing EHR data in predicting the four most common cancers. As shown in Table 6, other ML cancer papers mostly experimented with microarrays/expression data, imaging data, or publicly available datasets.

EHRs are used in hospitals to follow patients' conditions and histories. Physicians use them to keep track of patients' demographic information, diseases found during patients' hospitalization, diagnosis, therapies, lab results, etc. Frequent updates with new patient information make EHR databases sequential and information-rich. However, EHRs are intended for supervising hospital patients, but not for the research purposes like gene expressions, microarrays, and readily available datasets. Hence, it is significantly more difficult to handle data extraction and preprocessing for such data. Equally important, as Table 6 showed, the average medical dataset contains hundreds of examples, while we have tens of thousands, thus, making it even more challenging to find a suitable approach. We demonstrated that sequential ML models and data reduction methods could be a great combination in learning from sparse and high-dimensional data. Additionally, we presented that we cannot improve predictions even with data reduction methods if the sequence of patient visits is too long. Finally, the results of four classification tasks on a substantially large dataset support our claims and hypotheses.

Our study has several limitations. First, as we already mentioned, EHR databases are not utilized to collect information for research. Therefore, datasets of this type are often sparse and noisy, making them difficult to examine closely. Furthermore, we did not have access to all medical features from the records during the examination of the HCUP SID database. In other words, we could not take advantage of helpful information such as medical notes, lab results, and the first notice of "cancer in situ" (formation of abnormal cells). Finally, RNNs were only able to learn the order of patient visits, but not the time between the visits. Thus, this will be our future work. Suppose we manage to modify the cells of RNN models to consider the time between two consecutive visits of the patient. In that case, we could significantly increase

Table 6
Study comparisons based on the dataset type, size, predicted disease, models, prediction accuracy, and AUROC score. We compared our approach with the studies mentioned in Introduction and Related work sections.

Papers	Dataset & disease	Size	Model	Accuracy	AUROC
Our approach	EHR, four cancers	50,606	RNN GRU	81%–88%	0.88–0.9
Asri et al. [6]; 2016	WBCD, breast cancer	699	SVM	97.13%	–
Xie et al. [9]; 2021	Biomarkers, lung cancer	153	Naïve Bayes	100%	1
Huang et al. [10]; 2018	Gene expression, multiple cancers	175	SVM	82.6%	–
Pati [11]; 2019	Gene expression, lung cancer	96	MLP	86.7%	–
Wang et al. [12]; 2019	MRI, liver cancer	494	CNN	92%	–
Kadir and Gleeson [13]; 2018	CT images, lung cancer	1397	CNN	–	0.87
Liu et al. [14]; 2020	Laboratory data, liver cancer	2890	ANN	–	0.86–0.88
Zhang et al. [15]; 2020	Microarray data, liver cancer	1333	SVM	96.6%	–
Khourdifi and Bahaj [16]; 2018	WBCD, breast cancer	699	SVM	97.9%	–
Weegar and Sundström [24]; 2020	EHR, cervical caner	1.321	RF	–	0.9
Atrey et al. [25]; 2019	WBCD, breast cancer	699	ANN	99.6%	–
Li and Chen [26]; 2018	WBCD and BCCD, breast cancer	815	RF	74.3%–96.1%	0.785–0.989
Ferroni et al. [27]; 2019	Clinical and biochemical data, breast cancer	454	DSS	86%	–
Rajesh et al. [28]; 2020	UCI HCC survival, HCC	165	RF	80.64%	–
Wang et al. [29]; 2021	BioStudies database, HCC	535	RF	73.9%	0.803
Priya et al. [30]; 2018	UCI ILPD, liver cancer	583	J48 DT	69%–95%	–
Yuan et al. [31]; 2021	EHR, lung cancer	76,643	NLP & LR	–	92.7%
Miotto et al. [32]; 2016	EHR, liver cancer	700,000	DeepPatient	–	0.886
Wang et al. [33]; 2019	EHR, lung cancer	1000	MoEDL	–	0.75

prediction accuracy and use fewer hospitalizations to make the final prediction.

6. Conclusion

The results achieved in this study show that our approach, especially for RNN architectures, could predict the de-novo occurrence of cancer with high accuracy. GRU with an embedding layer could be potentially used as a decision support algorithm for early cancer detection by predicting hospitals' EHR data. When the cancer is detected early, patients have more treatment options and a far greater chance of survival.

Funding statement

This research was supported in part by the King Abdullah University of Science and Technology Center Partnership Fund and Pennsylvania Department of Health CURE Health Data Science Research Project.

Contributorship statement

JA prepared the datasets and selected adequate dimensionality reduction methods for this study. JA, BL, AAH, and MS wrote the source code. WD and MP preprocessed and participated in the preparation of the datasets. JA and BL designed the experimental setup and ran all experiments. Additionally, JA and BL developed the main idea behind the paper and received valuable feedback from ZO. All authors were involved in writing the paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors appreciate the Agency for Healthcare Research and Quality's Healthcare Cost and Utilization Project (HCUP) for supplying the data for this study.

References

[1] World Health Organization. accessed 17 January2022, https://www.who.int/.
[2] World Cancer Research Fund WCRF. 17 January2022, https://www.wcrf.org/.
[3] Key statistics about liver cancer. 17 January2022, https://www.cancer.org/cancer/liver-cancer/about/what-is-key-statistics.html.
[4] Ljubic B, Hai AA, Stanojevic M, et al. Predicting complications of diabetes mellitus using advanced machine learning algorithms. J Am Med Inf Assoc 2020;27:1343–51. https://doi.org/10.1093/jamia/ocaa120.
[5] Ljubic B, Roychoudhury S, Cao XH, et al. Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction. Comput Methods Progr Biomed 2020;197:105765. https://doi.org/10.1016/j.cmpb.2020.105765.
[6] Asri H, Mousannif H, Moatassime HA, et al. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput Sci 2016;83:1064–9. https://doi.org/10.1016/j.procs.2016.04.224.
[7] Shetty A, Shah V. Survey of cervical cancer prediction using machine learning: a comparative approach. In: 2018 9th International Conference on Computing, Communication and networking Technologies (ICCCNT). Published Online First; 2018. https://doi.org/10.1109/icccnt.2018.8494169.
[8] Louro J, Posso M, Hilton Boon M, et al. A systematic review and quality assessment of individualised breast cancer risk prediction models. Br J Cancer 2019;121:76–85. https://doi.org/10.1038/s41416-019-0476-8.
[9] Xie Y, Meng W-Y, Li R-Z, et al. Early lung cancer diagnostic biomarker discovery by machine learning methods. Transl Oncol 2021;14:100907. https://doi.org/10.1016/j.tranon.2020.100907.
[10] Huang C, Clayton EA, Matyunina LV, et al. Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. Sci Rep 2018;8. https://doi.org/10.1038/s41598-018-34753-5.
[11] Pati J. Gene expression analysis for early lung cancer prediction using Machine Learning Techniques: an eco-genomics approach. IEEE Access 2019;7:4232–8. https://doi.org/10.1109/access.2018.2886604.
[12] Wang CJ, Hamm CA, Savic LJ, et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 2019;29:3348–57. https://doi.org/10.1007/s00330-019-06214-8.
[13] Kadir T, Gleeson F. Lung cancer prediction using machine learning and Advanced Imaging Techniques. Transl Lung Cancer Res 2018;7:304–12. https://doi.org/10.21037/tlcr.2018.05.15.
[14] Liu X, Hou Y, Wang X, et al. Machine learning-based development and validation of a scoring system for progression-free survival in liver cancer. Hepatol Int 2020;14:567–76. https://doi.org/10.1007/s12072-020-10046-w.
[15] Zhang Z-M, Tan J-X, Wang F, et al. Early diagnosis of hepatocellular carcinoma using machine learning method. Front Bioeng Biotechnol 2020;8. https://doi.org/10.3389/fbioe.2020.00254.
[16] Khourdifi Y, Bahaj M. Applying best machine learning algorithms for breast cancer prediction and classification. In: 2018 International Conference on electronics, Control, optimization and Computer science (ICECOCS). Published Online First; 2018. https://doi.org/10.1109/icecocs.2018.8610632.
[17] HCUP. SID database Documentation. https://www.hcup-us.ahrq.gov/db/state/siddbdocumentation.jsp. 20 May2021.
[18] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
[19] Cho K, van Merrienboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on empirical methods in Natural Language processing (EMNLP). Published Online First; 2014. https://doi.org/10.3115/v1/d14-1179.
[20] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323:533–6. https://doi.org/10.1038/323533a0.
[21] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[22] Breiman L, Friedman JH, Olshen RA, et al. Classification and regression trees. Biometrics 1984;40:874. https://doi.org/10.2307/2530946.
[23] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. Am Statistician 1992;46:175. https://doi.org/10.2307/2685209.
[24] Weegar R, Sundström K. Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations. PLoS One 2020;15. https://doi.org/10.1371/journal.pone.0237911.
[25] Atrey K, Sharma Y, Bodhey NK, et al. Breast cancer prediction using dominance-based feature filtering approach: a Comparative Investigation in Machine Learning Archetype. Braz Arch Biol Technol 2019;62. https://doi.org/10.1590/1678-4324-2019180486.
[26] Li Y, Chen Z. Performance evaluation of machine learning methods for breast cancer prediction. Appl Comput Math 2018;7:212. https://doi.org/10.11648/j.acm.20180704.15.
[27] Ferroni P, Zanzotto F, Riondino S, et al. Breast cancer prognosis using a machine learning approach. Cancers 2019;11:328. https://doi.org/10.3390/cancers11030328.
[28] Rajesh S, Choudhury NA, Moulik S. Hepatocellular carcinoma (HCC) liver cancer prediction using machine learning algorithms. In: 2020 IEEE 17th India Council International Conference (INDICON). Published Online First; 2020. https://doi.org/10.1109/indicon49873.2020.9342443.
[29] Wang Y, Ji C, Wang Y, et al. Predicting postoperative liver cancer death outcomes with machine learning. Curr Med Res Opin 2021;37:629–34. https://doi.org/10.1080/03007995.2021.1885361.
[30] Priya MB, Juliet PL, Tamilselvi PR. Performance analysis of liver disease prediction using machine learning algorithms. Int Res J Eng Technol 2018;5:206–11.
[31] Yuan Q, Cai T, Hong C, et al. Performance of a machine learning algorithm using electronic health record data to identify and estimate survival in a longitudinal cohort of patients with lung cancer. JAMA Netw Open 2021;4. https://doi.org/10.1001/jamanetworkopen.2021.14723.
[32] Miotto R, Li L, Kidd BA, et al. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016;6. https://doi.org/10.1038/srep26094.
[33] Wang R, Weng Y, Zhou Z, et al. Multi-objective ensemble deep learning using electronic health records to predict outcomes after lung cancer radiotherapy. Phys Med Biol 2019;64:245005. https://doi.org/10.1088/1361-6560/ab555e.
[34] Choi E, Xiao C, Stewart WF, et al. Mime: multilevel medical embedding of electronic health records for predictive healthcare. 2018. arXiv preprint arXiv:1810.09593.
[35] Yu K, Zhang M, Cui T, et al. Monitoring icu mortality risk with a long short-term memory recurrent neural network. Biocomputing 2019;2020. https://doi.org/10.1142/9789811215636_0010. Published Online First.
[36] Li CH, Park SC. An efficient document classification model using an improved back propagation neural network and singular value decomposition. Expert Syst Appl 2009;36:3208–15. https://doi.org/10.1016/j.eswa.2008.01.014.
[37] Vo N, Hays J. Generalization in metric learning: should the embedding layer be embedding layer?. In: 2019 IEEE Winter Conference on applications of Computer vision (WACV). Published Online First; 2019. https://doi.org/10.1109/wacv.2019.00068.
[38] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.
[39] Chung J, Gulcehre C, Cho K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling Neural information processing systems (NIPS) Workshop on deep learning. 2014. arXiv preprint arXiv:1412.3555.

[40] Kingma DP, Ba J. ADAM: a method for stochastic optimization. In: International Conference on learning representations (ICLR); 2015. arXiv preprint arXiv: 1412.6980.

[41] Masters D, Luschi C. Revisiting small batch training for deep neural networks. 2018, 07612. arXiv preprint arXiv:1804.

[42] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30. https://doi.org/10.5555/1953048.2.