Privacy Preserving Survival Prediction

Stefano Fedeli
KTH Royal Institute of Technology
Stockholm, Sweden
fedeli@kth.se

Frida Schain
Schain Research
Stockholm, Sweden
frida@schainresearch.com

Sana Imtiaz
KTH Royal Institute of Technology
Stockholm, Sweden
sanaim@kth.se

Zainab Abbas
KTH Royal Institute of Technology
Stockholm, Sweden
zainabab@kth.se

Vladimir Vlassov
KTH Royal Institute of Technology
Stockholm, Sweden
vladv@kth.se

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9672036

***Abstract*—Predictive modeling has the potential to improve risk stratification of cancer patients and thereby contribute to optimized treatment strategies and better outcomes for patients in clinical practice. To develop robust predictive models for decision-making in healthcare, sensitive patient-level data is often required when developing the training models. Consequently, data privacy is an important aspect to consider when building these predictive models and in subsequent communication of the results. In this study we have used Graph Neural Networks for survival prediction, and compared the accuracy to state-of-the-art prediction models after applying Differential Privacy and k-Anonymity, i.e. two privacy-preservation solutions. By using two different data sources we demonstrated that Graph Neural Networks and Survival Forests are the two most well-performing survival prediction methods when used in combination with privacy preservation solutions. Furthermore, when the predictive model was built using clinical expertise in the specific area of interest, the prediction accuracy of the proposed knowledge based graph model drops by at most 10% when used with privacy preservation solutions. Our proposed knowledge based graph is therefore more suitable to be used in combination with privacy preservation solutions as compared to other graph models.**

***Index Terms*—knowledge graph, survival prediction, privacy preservation, differential privacy, anonymization, clinical data, national registry, graph neural network, survival forest**

I. Introduction

In recent years there has been a strong trend to use real-world clinical and genetic data to guide healthcare decisions and enable personalized treatment. This concept is usually referred to as personalized medicine or precision medicine. Artificial intelligence (AI), genomics, and identified digital biomarkers, in combination with the digitalization of healthcare, have been important drivers in this development. Previous studies have described the role of machine learning in diagnostic prediction [1]. Similarly, machine learning (ML) has also been used to develop survival prediction models in the field of cancer research [2]. Random forests in particular have been used by many groups to study survival prediction. Moreover, Graph Neural Network (GNN) is another interesting method given the high accuracy demonstrated in cancer survival prediction [3] [4].

Graphs are useful for representing relational data in various application domains, such as social media analysis [5], web analysis [5], recommendation systems [6], knowledge graphs [7] and road traffic analysis [8]. Recently, graph embeddings (node/vertex embeddings) are gaining popularity due to their capabilities of capturing graph structure and node information for various homogenous or heterogenous knowledge graphs [9], [10]. These embeddings are then fed to a neural network that can perform downstream tasks, e.g., node classification, clustering and link predictions. Previous works have shown promising results in the field of survival prediction using graphs to model genetic data [3] [4]. In this work, we propose to use graph embeddings and graph neural networks to learn the knowledge based graph of patient data without using genetic data and do survival prediction while preserving the privacy of sensitive data.

Despite the potential benefits associated with predictive modeling to guide treatment decisions in clinical practice, there are some challenges to be taken into consideration. For example, the predictive methods used are not yet widely accepted, partly due to data privacy concerns, lack of generalization and unavailability of structured data [11]. Data privacy aspects related to sensitive patient data are of particular importance when working with training and making inferences from ML models. In this study, we used patient-level data from the Swedish National Board of Health and Welfare to evaluate how ML models behave when privacy-preserving techniques are applied. To generalize our findings, we performed the experiments with multiple ML models and a benchmark dataset. We evaluated all the models on the area under the curve (AUC) scores on a binary classification task.

The contributions of this work are as follows.

- We propose to represent cancer datasets as graphs representing useful relations between the patients as nodes and characteristics as edges of the medical knowledge graph.
- We propose to use Graph Neural Networks to perform survival prediction for real-world cancer patients using GraphSAGE, a popular graph representation learning framework. Our model leverages relations between patients which are useful for survival prediction.

- We evaluated the impact of using privacy-preserving solutions, namely k-Anonymity and Differential Privacy, on the predictive performance of state-of-the-art survival prediction techniques compared with our graph-based prediction techniques.
- We evaluated the trade-off between prediction accuracy and privacy preservation for our graph-based survival prediction method and state-of-the-art survival prediction techniques for sensitive datasets.

The rest of this paper is structured as follows. Section II gives an overview of the related background and state-of-the-art regarding both survival prediction and privacy preservation. Section III provides information about our choice of the data and the chosen use case for our experiments. Section IV describes the survival prediction models used for our experiments. Section V presents our experimental setup and results. We also provide reasoning and in-depth discussion on our obtained results, trying to connect the dots and propose explanations. We present our conclusions, identify the limitations of our work and present the analyses of the future directions we foresee in Section VI.

II. Background and Related Work

Survival Prediction. Survival prediction is a development of traditional survival analysis, a branch of statistics focusing on modeling the problem of time-to-event [12]. Historically, regression analysis has been used to estimate relationships between certain variables and outcomes of interest (e.g., survival). This approach however has not been pursued in survival prediction due to poor performances, co-founding factors and lack of data collected [13].

Non-linear models such as survival forests have been associated with promising results when using risk ranking for survival prediction. Data is divided in each layer into two groups that show the minimal p-values in the log-rank test between their two Kaplan Maier curves [14]. The survival forest method was developed from random forests that are also commonly used models in survival prediction [15], mostly when the problem is shaped as a binary classification task. Binary classification using disease-specific thresholds is also commonly used in machine learning for survival prediction.

These models are all associated with different limitations which are particularly evident when working with data including many dimensions [16]. Genetic aberrations are important predictors for survival in many cancer types such as lung cancer and breast cancer [17]. Genetic data may be complex with many dimensions, and previous studies have shown that graph-based models perform better in these settings compared to, for instance, regression models [3] [4]. Other groups have compared different predictive models in terms of accuracy performances, but to the best of our knowledge no study assessed different models using privacy-preserved real-world data [18], [19].

Graph Representation. Graphs can model complex and multi-dimensional data and is therefore suitable for healthcare data. Multimodal cellular networks are commonly used in biology research and drug discovery, but also related to transactions, social networks, and behaviors on the internet [20]. A network can be modeled as a graph G, a tuple containing two sets: the first one consists of vertices V that represent the entities of the graph, also called nodes; the second one consists of edges E, also called links that connect vertices.

$$G = (V, E) \tag{1}$$

Compared to other data structures, graphs put the focus on relationships between nodes instead of their property. The strength of graphs becomes a weakness when such data structure is used to feed a machine-learning algorithm. For this reason, many techniques have been developed to perform feature engineering against a graph structure and provide a better Euclidean representation that could be used by the machine learning method. Starting with graphlets or triangles counts, many were the common approaches to extract useful information from graph relationships [21] [22]. While interesting and useful to some extent, those methods can capture only a small portion of all the information that a graph or a node is carrying. Those methods are limited by our comprehension of graphs and that is the main reason a never-ending shift to automatic representation is flourishing.

Graph Representation Learning can be seen as a simple function that can bring a graph, a node, or an edge from a sparse, non-euclidean space to a point in a latent space of m dimension. The m coordinates of this point in this latent space is called embedding of x where x is the argument of the mapping function:

$$y = f(x), y \in R^m \tag{2}$$

Graph Neural Networks. Graph Neural Networks (GNNs) are meant to be able to capture the right features of a complex data structure such as, in this case, a graph.

GraphSAGE is one of the most commonly used GNNs that learn to create embeddings by aggregating information from a sampled neighborhood [10]. GraphSAGE is working under the assumption that nodes that reside in the same neighborhood should have similar embeddings. For this reason, it exists a parameter K that encodes the maximum hop-distance node that can influence a node embedding.

The way GraphSAGE aggregate this information is customizable based on the task. The authors proposed three aggregation functions that can be relevant in many different tasks but for our work, we will focus on the most naïve and simple aggregator: the MEAN. In this case, the aggregation is simply based on computing the element-wise mean of the incoming feature vectors of the single nodes as shown below:

$$h_N^l(v) \leftarrow MEAN(\{h_u^{l-1}, \forall u \in N(v)\}) \tag{3}$$

Privacy Preservation. We can spot many publications that take into account all those different predictive models and compare them in terms of accuracy performances, but nothing has been said about their applicability in the

real world [18], [19]. In clinical practice, data is often processed in an anonymized way and it is very difficult to deploy an application that uses unprotected data. This applies also to researchers in countries, such as Sweden, where the data available is often available but truncated or aggregated, giving little or no room to achieve good results with the models we have encountered in the literature. It is, therefore, crucial to understand how all the models that we mentioned before behave when the data has been privatized by the source to guarantee the safety of all the actors involved.

k-Anonymity. To the best of our knowledge, k-anonymity is the most common technique used broadly to protect clinical records. It is the standard named by the 'Family Educational Rights and Privacy Act (FERPA)' of the US and the 'Guidelines for De-identification of Personal Data' of South Korea [23]. k-anonymity is often the best choice as it is easy to implement and is also efficient because it does not create unnecessary computational overhead. k-anonymity is based on the concept of aggregation and suppression. A dataset is defined as k-anonymous if each record contained in a released dataset cannot be distinguished from at least $k-1$ other individuals. This can be achieved by aggregating information in classes or removing some critical information, keeping an eye on avoiding over-generalization. It is possible to notice that the higher is k the more complex would be the task for the model but indeed it is still possible to leak some useful information. k-anonymity is part of the anonymization technique family that guarantees the protection of personally identifiable information by the removal of sensitive attributes such as ID, name, age or gender, or race. Unfortunately, anonymization techniques are susceptible to attribute disclosure attacks as well as database reconstruction attacks [24]. Attribute disclosure attacks are the most difficult to contrast as they consist in realizing some information that could be used to infer sensitive information by itself or when linked to other information coming from the same or different source dataset. Against such kind of attack, differential privacy is often used and that's the reason why, for our work, we focus on this latter privacy-preserving method.

Differential Privacy. Differential Privacy (DP) is a rigorous mathematical framework that defines an algorithm to be differentially private if and only if the inclusion of a single new record in the dataset causes only a limited, non statistically significant, change in the output of the function f. More formally Differential Privacy is defined as [25]:

"A randomized mechanism K provides (ϵ, δ)-differential privacy, if for any two neighboring database D_1 and D_2 that differ in only a single entry, $\forall S \in Range(K)$,"

$$Pr(K(D_1) \in S) \leq e^{\epsilon} Pr(K(D_2) \in S) + \delta \qquad (4)$$

If $\delta = 0$ then K is said to be ϵ-differentially private. With a differentially private algorithm, we can add noise to the function f and guarantee privatized results. This noise is proportional to the sensitivity of the output which means the maximum amount of output's change caused by the insertion of a single instance. The most popular mechanism, and the one we will use in our experiments, to achieve differential privacy on a dataset D, is a Laplacian noise [26].

A. Related Work

As mentioned in Section I, we focus on the binary classification task which is the most common approach to survival prediction in the ML literature.

Lynch et al. [27] in their work discussed how machine learning techniques can be used to predict survival, confirming results from another similar work by Walczak and Velanovich [28]. This work is based on using a simple neural network model that acts as a feed-forward network on two layers. This architecture was able to perform as well as the COX baseline [29] in the task of predicting if an individual will survive more or less than 7 months.

In 2019, Daoud and Mayo [19] published a work that showed at least ten neural network architectures have been used to handle genomic data and predict individual survival with very interesting results. Regarding graphs, Wang et al. [4] have proposed a complex architecture that performs convolutions on a graph where nodes are individuals and edges are weighted based on individual similarity among m dimensions. A strong emphasis is put on data preprocessing, which tries to build first a similarity network for each of the m dimensions and then merge all in a single adjacency matrix. Once the adjacency matrix is built, every node, that represents a patient, is assigned to a feature vector that will be used to create their embedding through a series of convolutions. A similar approach to graph modeling is the one proposed by Gao et al. [3] in their paper published in 2020. The architecture tries to merge different information coming from different sources by modeling each of them in the best possible way. In their architecture graphs are used to model the relationships between individuals and gene activations creating a bipartite graph where nodes are either of type patient or type gene. The graph embedding concatenated to the standardized node feature vector is given in input to a neural network that performs the task.

Both DP and k-anonymity have already a big track of records regarding their applicability and performances in the different scenarios but there is a gap when we are talking about GNN and graphs models in general. Wimmer et al. [30] have tested k-anonymity against a set of well-known machine learning models using three different benchmark datasets, finding that in certain cases anonymization could lead to an increase of performances of the models. This is due to a reduction of overfitting that k-anonymity brings by construction. Abadi et al. [31] have instead proposed an efficient way to apply DP to neural networks to avoid a big loss of performance. Their findings could be extended to other machine learning models proving that DP can be applied at a manageable cost. The work done by Johansson et al. addresses DP on graphs classification [32] is also very interesting. They show the efficacy in the

trade-off between accuracy and privacy on graph classification by developing a particular DP version of random walks and graphlets.

III. Data

To properly test the resilience of survival prediction models including privacy-preserving techniques, it is critical to have rich data from the real world. In this section we explain various real world cancer datasets used in our work of survival prediction.

A. Data Sources

In this study, we used data collected from the following sources.

- Swedish population-based national register data provided by the National Board of Health and Welfare after ethical approval by the Swedish Ethical Board (DNR 2021-07259). The study population included 3691 unique patients that underwent hematopoietic stem cell transplantation (HSCT) due to hematological malignancy. Patients were followed longitudinally, and the median follow-up time was 7,74 years. The follow-up time feature was used to label the records for our binary classification task using the threshold of three years. Ten clinically relevant variables with acceptable data completeness were selected for the model as shown in Table I.
- The TCGA-BRCA repository [1] from the NCI's Genomic Data Commons (GDC) was analyzed separately and used as a benchmark. This dataset included data from 1098 patients diagnosed with breast cancer with a median followup time of 3,20 years. TCGA-BRCA data source is commonly used by the cancer research community with a unified data repository that enables data sharing across cancer genomic studies in support of precision medicine.

B. Datasets

The Swedish data (defined as V1) contained 37% patients that survived less than 3 years after the HSCT procedure. To build this dataset, we picked 10 rich features of 3691 patients that do not include skewed or missing values from the many available features.

The GDC TCGA-BRCA dataset (defined as V2) is common for benchmarking new algorithms that focuses on genes, for instance, it was used by Wang et al. [4] to evaluate their architecture. To enable an indirect comparison to the results published by Wang et al. [4], a similar preprocessing pipeline was used. The dataset includes 1066 individuals with 20 features, shown in Table II.

In both cases, the training was adjusted to deal with unbalanced labels.

TABLE I
V1 Features

Feature	Type
Sex	boolean
Age at index (years)	numeric
Transplant calendar year	numeric
Comorbidity Index	numeric
Duration of the transplant (days)	numeric
Ratio of time spent in the hospital prior transplant (%)	numeric
Hospital	categorical
Cancer Diagnosis	categorical
Cancer Group	categorical
Type of Transplant	categorical

TABLE II
V2 Features

Feature	Type
Age at index (years)	numeric
Ethnicity	categorical
Race	categorical
Age at diagnosis (years)	numeric
M Stage	categorical
N Stage	categorical
T Stage	categorical
Stage At Diagnosis	categorical
Staging System	categorical
Diagnosis ICD10	categorical
Morphology	categorical
Primary Diagnosis	categorical
Prior Malignancy	boolean
Prior Treatment	boolean
Site of Biopsy	categorical
Synchronous Malignancy	categorical
Tissue of Origin	categorical
Therapy	boolean
Treatment Type	categorical
Diagnosis calendar year	numeric

IV. Survival Prediction Models

Based on our findings in the literature we decided to use two ensemble methods, one artificial neural network and one GNN. One of the two ensemble methods, the survival forest, served as a baseline for comparative purposes. Given the nature of the survival forest, this model has been adapted to binary classify patients instead of ranking them. This was accomplished by comparing the ranking provided by the forest (Fig.1 left) with the real known ranking (Fig.1 right) that is computed by ordering the patient based on their death date. This enabled us to find all the components of the confusion matrix, as shown in Figure 1, that will be later used to compute the metrics in the experiment.

The other ensemble model was a Random Forest model implemented with scikit [2] as it is the most common traditional ML model. Random Forest was also used by Gao et al. [3] in their work against the TCGA-BRCA benchmark.

Lastly, we also use a Deep neural network, denoted as DL/ANN in our experiments, for comparison with all the baselines methods and our graph-based prediction models to

[1] https://portal.gdc.cancer.gov/projects/TCGA-BRCA

[2] https://scikit-learn.org/stable/index.html

TABLE III
DISTRIBUTION ACROSS DIAGNOSIS IN V1 DATASET

Diagnosis Group	ICD10	Number of Patients
Lymphocytic Leukemia	C91.*	608
Myeloid Leukemia	C92.*	1260
Multiple Myeloma	C90.*	311
Other Leukemia	C9*	113
Malignant Neoplasms	C8*	405
Other Malignant Cancers	C*	510
Blood Diseases	D*	312
Other Diseases	*	173

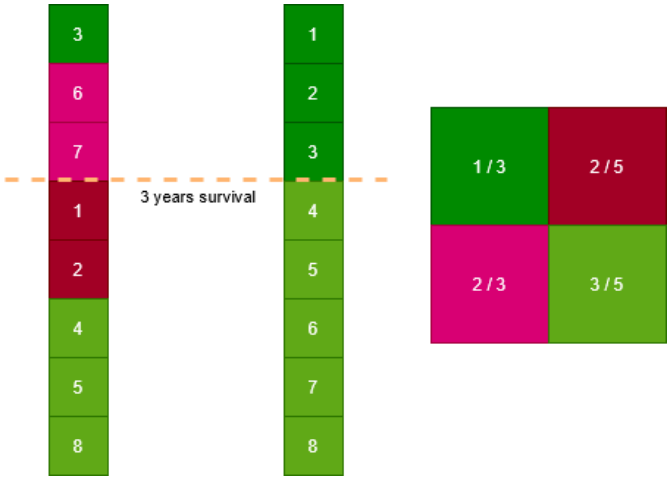

Fig. 1. Survival Forest as a classifier

evaluate how is the simple neural network going to perform compared to complex prediction models.

Recent work demonstrated impressive performances by applying GNNs to the problem of survival prediction [3] [4]. Although genetic data was modeled as graphs in the aforementioned works. It is interesting to shape our tabular data as a graph and apply recent GNN algorithms. Next, we explain how we modelled our tabular data of cancer patients as graphs.

We created three different types of graph models, namely G1, G2, and G3.

a) HinSage (G1): The graph model G1 is based on the Swedish dataset V1 using the knowledge graph shown in Figure 2. The blue nodes in the figure represent patients and orange nodes represent cancer types. Patient's information, shown in the patient table in Figure 2, is used for creating node embeddings. Similarly, the cancer node also contains cancer information that is used to create embeddings for cancer nodes. G1 is a heterogeneous bipartite graph where we have two sets of nodes, i.e., patient and cancer types.

The complete architecture for survival prediction using G1 is shown in Figure 3. In the first step, the heterogeneous graph G1 is created and fed to the graph embedding module, which generates low-dimensional graph embeddings. These embeddings are then fed to the fully connected neural network that does survival prediction. The graph embedding module uses HinSage for G1. HinSage is a variation of GraphSage that takes into account heterogeneous nodes by assigning different weights to edges depending upon the relationship.

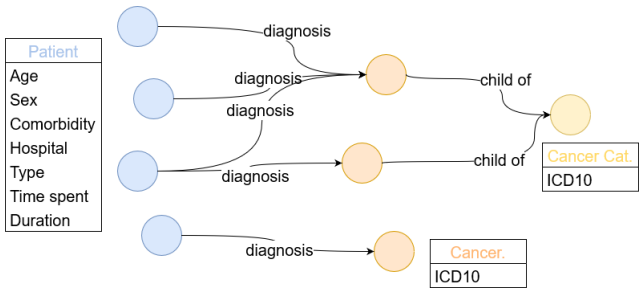

Fig. 2. Graph Model used by G1 for Swedish dataset V1

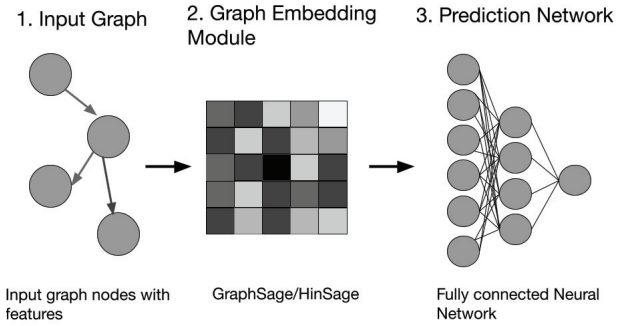

Fig. 3. Graph Based Survival Prediction Model Architecture

b) GraphSage/snf (G2): G2 is built using the similarity network fusion concept from Wang et al. [4]. The similarity value is useful to connect the nodes of the graph. The idea is to create many different similarity networks on different dimensions and then fuse them using the Algorithm 1.

Algorithm 1: Similarity Network Fusion (SNF)

Input: m Exponential similarity matrices $W(i,j)$
Output: Similarity Matrix $P^{(c)}$
$S(i,j) \leftarrow \frac{W(i,j)}{\sum_{k \in N_i} W(i,k)}$
for z = *1...m* **do**
  $P^z(i,j) \leftarrow \frac{W(i,j)}{2\sum_{k \neq i} W(i,k)}$
$P^{(c)} \leftarrow \frac{\sum_{k=1..m} P^{(k)}}{m}$

In the case of the dataset V2, we set the parameter $m = 1$ for the similarity algorithm, which means that no fusion was done and just the similarity was computed. This is due to the fact that the data was collected clinically and no external data source was used. Figure 4 shows the pre-processing of the data, where the similarity between two sets of clinical data is being computed to generate a homogeneous graph G2 which is later used for survival prediction using the similar architecture as shown in Figure 3. For G2, we use GraphSage in the embedding module because all edges are treated equally since G2 is a homogeneous graph. We term this graph as

homogeneous because all nodes are of the same type, i.e., patient nodes containing all the patient attributes along with their cancer type information.

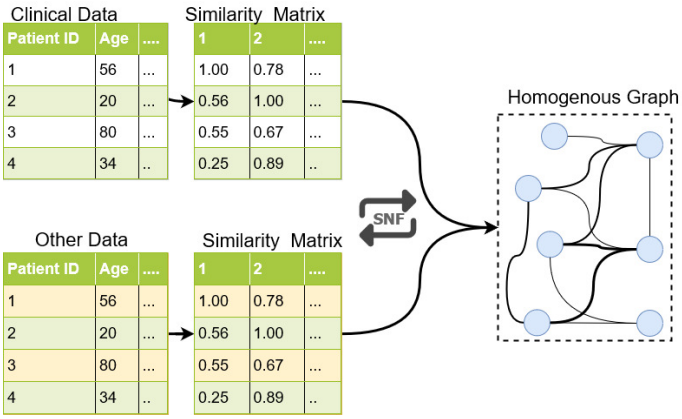

Fig. 4. G2 Similarity Network Fusion

c) GraphSage/cosine (G3): The last graph model G3 used in the experiments is a simplified version of the graph used in [4]. G3 is built from a similarity matrix $N \times N$ where N is the number of individuals in the dataset. The similarity matrix is built using the cosine similarity which is the most effective metric when using numerical variables. We use a simple cosine similarity between the row of each dataset to improve the performance compared to [4], where different data sources are used in similarity computation. A threshold of 0.7 has been used to draw the only relevant connections between nodes to make sure that we do not end up with a fully connected graph.

Figure 5 shows the process to build the graph, where the similarity between rows of the data is being computed. If the similarity is high then an edge is created between the patient nodes of the corresponding rows. This homogeneous graph G3 is then fed to a GNN which learns the graph embeddings using GraphSage and feeds them to the NN for survival prediction as shown in Figure 3.

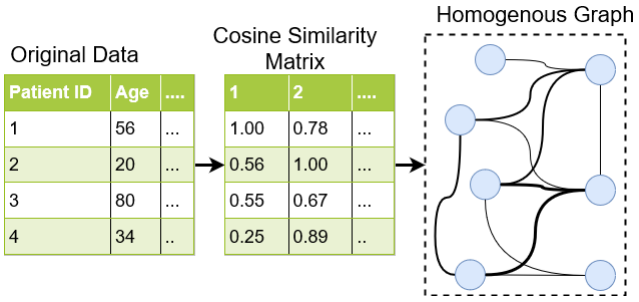

Fig. 5. G3 Building process

We categorize the aforementioned graph models into two groups.

- **Knowledge-Driven architectures (G1)**: In this architecture, the input of domain experts is used to create the graph. Edges are created under known and relevant relationships. For the Swedish dataset V1, relationships are created between patients and diagnosis, carefully designing the nodes and the connection between them.
- **Data-Driven architectures (G2, G3)**: Are all architectures are fully based on data. The graph here is built by considering each record of the dataset as a node and then connecting them using different notions of similarity, such as Figure 5, and network fusion, such as Figure 4, [4].

Regarding Graph Neural Networks algorithms, we have chosen to experiment with GraphSage and its heterogeneous version HinSage [10]. The three different graph models are used as input for the GNN in our experiment. This choice takes into high consideration the real-world application that this algorithm could have in the future. It is also very important to validate how those algorithms will be impacted by privacy techniques due to the fact they will be likely to be used in dynamic and real scenario settings.

V. Evaluation

We performed two distinct experiments. One with the datasets privatized with k-anonymity and the other with the datasets privatized using Differential Privacy. Both algorithms had one hyperparameter to increase or decrease the privacy injected into the data. This led us to perform the two experiments with different values of the hyperparameters and evaluate how these influenced the results.

A. Experimental Setup

For the parameter k in k-anonymity we verified for $k \in \{3, 9, 30\}$ while for the parameter ϵ of Differential Privacy we try for $\epsilon \in \{2, 1, 0.5, 0.1\}$. This choice was driven by using the most common values found in the literature.

We aimed to protect selected attributes of each record. However, we decided to anonymize the dataset over all the attributes when using k-anonymity while using only a subset of attributes when applying differential privacy.

We split each dataset in test (20%) and train (80%) using then cross-validation to grid-searching the models' hyperparameters. This allowed us to be sure we were running the different models at their probably best.

To evaluate and compare the results of the different scenarios and models, the study focused on quantitative metrics to assess the differences. Following previous work published [3], [33], [34] in the area, it is clear that the most important metrics one should consider are AUC, Recall, and Precision. When facing a binary classification problem the most important quantitative value to be considered is the combination of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

After running each experiment 5 times we computed the average as it is shown by the results reported in Tables IV, V, VI, and VII.

TABLE IV
AVERAGE AUC AFTER 5 EXPERIMENTS WITH DIFFERENT K ON V1

Models	V1			
	k=1	k=3	k=9	k=30
Survival Forest (SF)	0.659	0.647	0.641	0.630
Random Forest (RF)	0.605	0.594	0.580	0.573
DL/ANN	0.660	0.622	0.600	0.613
HinSage (G1)	0.662	0.628	0.626	0.611
GraphSage/snf (G2)	0.644	0.643	0.620	0.612
GraphSage/cosine (G3)	0.660	0.685	0.630	0.638

TABLE V
AVERAGE AUC AFTER 5 EXPERIMENTS WITH DIFFERENT K ON V2

Models	V2			
	k=1	k=3	k=9	k=30
Survival Forest (SF)	0.703	0.688	0.662	0.599
Random Forest (RF)	0.818	0.813	0.793	0.795
DL/ANN	0.824	0.793	0.773	0.768
HinSage (G1)	0.845	0.840	0.816	0.793
GraphSage/snf (G2)	0.857	0.775	0.731	0.716
GraphSage/cosine (G3)	0.803	0.800	0.731	0.717

Regardless of which technique or model that was used, a higher degree of privacy-preservation was always associated with lower performance. The performance decline was similar for all models, except for survival forest and G1, as shown by both Figure 6, 7. Another general consideration is that even though the privatization increased ten times, the performances did not decline with the same proportion. What seems to be happening is that at some point performances reach a plateau, and after that, we observed a convergence to the performance of a random classifier. This was obvious with k-anonymity, Figure 6, where after $k = 9$ the performances did not follow anymore the trend but slowly drop to 0.5.

Furthermore the data showed that survival forests are very resistant to differential privacy and much less reliable when k-anonymity was applied. Figure 8 shows the variation of the AUC across the hyperparameter variations therefore models that are consistent across the different runs will show a short box. In chart (a) survival forest shows a very narrow box underlying its resilience to noise addiction. By contrast, in the chart (b), the survival forest shows high susceptibility to anonymization.

From Figures 6 and 7 we can observe a similar pattern where most of the model decrease their performances as privacy increased. However, those performances' reduction are related to the dataset as graphs' models did not show similar behavior in the two datasets even though the privacy technique is the same. In particular, looking at the results from V2, in Figure 7, most of the models lost around 10%-20% AUC compared to the 5% loss in AUC that we have, on average, on the V1 dataset. This particular behavior can be the product of the difference in feature size of the dataset. More features could indeed amplify the effect of privacy-preserving techniques.

TABLE VI
AVERAGE AUC AFTER 5 EXPERIMENTS WITH DIFFERENT ϵ ON V1

Models	V1				
	$\epsilon = \infty$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.1$
Survival Forest (SF)	0.659	0.631	0.616	0.609	0.607
Random Forest (RF)	0.605	0.608	0.603	0.579	0.570
DL/ANN	0.660	0.601	0.572	0.573	0.527
HinSage (G1)	0.662	0.635	0.633	0.629	0.609
GraphSage/snf (G2)	0.644	0.639	0.631	0.617	0.596
GraphSage/cosine (G3)	0.660	0.642	0.618	0.616	0.603

TABLE VII
AVERAGE AUC AFTER 5 EXPERIMENTS WITH DIFFERENT ϵ ON V2

Models	V2				
	$\epsilon = \infty$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.1$
Survival Forest (SF)	0.703	0.697	0.692	0.665	0.662
Random Forest (RF)	0.818	0.795	0.789	0.768	0.691
DL/ANN	0.824	0.742	0.738	0.723	0.693
HinSage (G1)	0.845	0.858	0.846	0.808	0.754
GraphSage/snf (G2)	0.857	0.801	0.769	0.767	0.641
GraphSage/cosine (G3)	0.803	0.800	0.772	0.771	0.673

The behavior of random forest is almost opposite to the other models. On average, it is the most resilient model to privacy preserving techniques we have tested. Their performances are reduced by -0.127 at maximum in term of AUC. Low, compared to the biggest loss of -0.216 by G2. Another behavior is the sudden drop of the random forest performance as the amount of privacy reaches critical values that correspond to a very high amount of noise.

Among the graph models, G1 is the one that is more resilient and never drops its AUC more than 10%. G2 instead is near 26%.

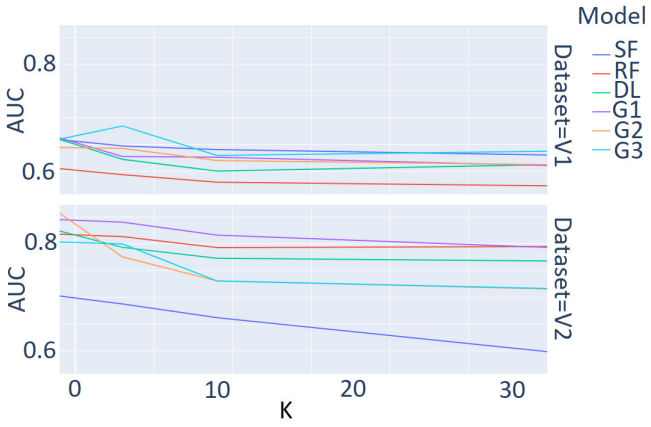

Fig. 6. Trendline chart as k varies across 6 models on V1 and V2

B. Discussion

The results from the present study indicated that G1 is the most resilient model to privacy-preserving techniques. Although the performance was impacted by the privacy-preserving techniques, it remains always the top performer

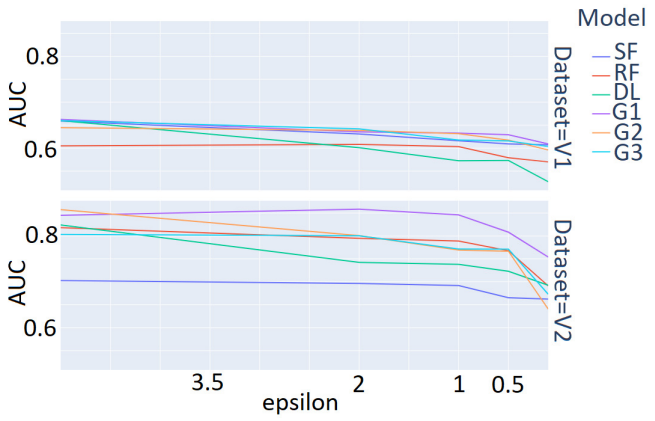

Fig. 7. Trendline chart as ϵ varies across 6 models on V1 and V2

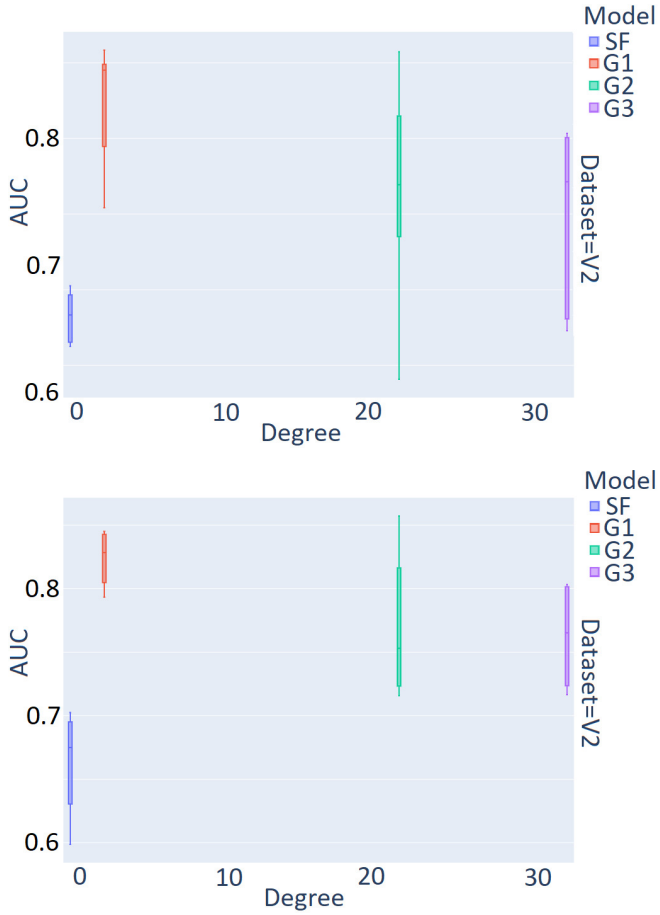

Fig. 8. Performance of graph models and baseline in relation to their average degree when (a) Differential Privacy (b) k-anonymity is applied.

in most of the experiments. As described in section IV, the dataset, in this case, is modeled as a heterogenous graph that has patients nodes and clinical insights nodes. The relationships we draw in the case of G1, opposite to G2 and G3, are few as shown by Figure 8 and very carefully selected. The reason for this high resiliency may be due to the few amount of features used to draw the relationship, e.g. 1, that gave the GNNs used, HinSage, the insight of which feature was the most important to focus on among the many available. With less relationship between nodes, differential privacy had less impact on the graph.

When applying k-anonymity we noticed that the graph models instead remained quite resilient while survival forest seems to be very dependent on the data, shown by Figure 6. The reason behind such behavior can be related to the relationship of the different nodes. k-anonymity is bringing many records to look the same and therefore traditional models struggle. Many records in the dataset look identical but might be labeled differently. Graph modeling, and therefore GNNs, which focus on the relationships, are still able to classify the records because they can leverage the node's relationships with the other nodes even though the node's feature vector looks the same as many others. In this way, graph neural networks do not perish a great impact from k-anonymity. On the opposite, ensemble methods are completely misguided by the privacy-preserving method as they assume each record is independent of the others.

The same reasoning cannot be applied for differential privacy as shown in Table VII. Laplacian noise-damaged all the models, despite survival forests and HinSage. As the noise reached high levels, such as 0.1 epsilon, we observed a reduced AUC. The GNNs poor performances can be related to intensively random changes in the graph structure, which would inhibit GNNs from learning. This was confirmed by comparing graph models based on similarity matrices with G2 that is instead created by leveraging clinical knowledge. G2 is more resilient to noise as relationships are based on a single feature instead of the whole dataset. On the contrary, survival forest appears more resilient due to its inner structure and the ability to rely on a ranking instead of a binary label when building the trees.

VI. Conclusions and Future Work

In this work, we have investigated how ML methods such as GNNs can be applied in prediction when privacy-preserving techniques are applied. Results showed that the difference in accuracy was highly dependent on the modeling of the data, e.g. how the graph was created, and less dependent on the ML architecture used. In particular, graph modeling which focus on drawing only crucial, clinically relevant, relationships is the most resilient model among the ones we have tested. In our experiments, we show that when data is modeled in this way, the most significant drop in performance was experienced when data is noised via differential privacy. In this case, the model performances dropped by 10% while in all the other cases we observed a lighter reduction. Clinical models, such as survival forests, have shown good behavior against noise or anonymization with respect to traditional ML models, such as a random forest.

In summary, this work demonstrated that graphs can be useful tools when representing complex real-world data in sur-

vival prediction models and that privacy-preserving techniques can be applied with acceptable performance.

A. Future Work

The work has some limitations concerning the privacy-preserving techniques since we focused on only two privacy-preserving methods among many different available in the literature. Results and conclusions can change if different privacy-preserving techniques are applied to GNNs. For this reason, future work should also explore how graphs are impacted by different types of privacy-preserving techniques. Working with survival prediction models, it is important to add the right clinical expertise when interpreting results derived from ML techniques. Ying et al. [35] recently presented a framework based on mutual independence of subgraphs to show which node features that have the highest impact on the prediction given by the network. Future studies are needed to further explore how these techniques can be useful in clinical practice.

REFERENCES

[1] M. Fatima, M. Pasha *et al.*, "Survey of machine learning algorithms for disease diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, p. 1, 2017.

[2] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer informatics*, vol. 2, p. 117693510600200030, 2006.

[3] J. Gao, T. Lyu, F. Xiong, J. Wang, W. Ke, and Z. Li, "Mgnn: A multimodal graph neural network for predicting the survival of cancer patients," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1697–1700.

[4] C. Wang, J. Guo, N. Zhao, Y. Liu, X. Liu, G. Liu, and M. Guo, "A cancer survival prediction method based on graph convolutional network," *IEEE transactions on nanobioscience*, vol. 19, no. 1, pp. 117–126, 2019.

[5] I. Stanton and G. Kliot, "Streaming graph partitioning for large distributed graphs," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 1222–1230. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339722

[6] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang, and G. de Melo, "Fairness-aware explainable recommendation over knowledge graphs," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 69–78. [Online]. Available: https://doi.org/10.1145/3397271.3401051

[7] Y. Cao, L. Huang, H. Ji, X. Chen, and J. Li, "Bridge text and knowledge by learning multi-prototype entity mention embedding," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1623–1633.

[8] Z. Abbas, J. R. Ivarsson, A. Al-Shishtawy, and V. Vlassov, "Scaling deep learning models for large spatial time-series forecasting," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 1587–1594.

[9] W. L. Hamilton, "Graph representation learning," *Synthesis Lectures on Artifical Intelligence and Machine Learning*, vol. 14, no. 3, pp. 1–159, 2020.

[10] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *arXiv preprint arXiv:1706.02216*, 2017.

[11] E. D. Peterson, "Machine learning, predictive analytics, and clinical practice: can the past inform the present?" *JAMA*, vol. 322, no. 23, pp. 2283–2284, 2019.

[12] C.-F. Chung, P. Schmidt, and A. D. Witte, "Survival analysis: A survey," *Journal of Quantitative Criminology*, vol. 7, no. 1, pp. 59–98, 1991.

[13] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[14] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[15] M. Zakariah *et al.*, "Classification of large datasets using random forest algorithm in various applications: Survey," *International Journal of Engineering and Innovative Technology (IJJEIT)*, vol. 4, no. 3, 2014.

[16] A. Dhillon and A. Singh, "Machine learning in healthcare data analysis: a survey," *Journal of Biology and Today's World*, vol. 8, no. 6, pp. 1–10, 2019.

[17] F. J. Couch, K. L. Nathanson, and K. Offit, "Two decades after brca: setting paradigms in personalized cancer care and prevention," *Science*, vol. 343, no. 6178, pp. 1466–1470, 2014.

[18] K. M. Mendez, S. N. Reinke, and D. I. Broadhurst, "A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification," *Metabolomics*, vol. 15, no. 12, pp. 1–15, 2019.

[19] M. Daoud and M. Mayo, "A survey of neural network-based cancer prediction models from microarray data," *Artificial intelligence in medicine*, vol. 97, pp. 204–214, 2019.

[20] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant *et al.*, "Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic acids research*, vol. 43, no. D1, pp. D1071–D1078, 2015.

[21] N. Pržulj, D. G. Corneil, and I. Jurisica, "Modeling interactome: scale-free or geometric?" *Bioinformatics*, vol. 20, no. 18, pp. 3508–3515, 2004.

[22] D. Marcus and Y. Shavitt, "Rage–a rapid graphlet enumerator for large networks," *Computer Networks*, vol. 56, no. 2, pp. 810–819, 2012.

[23] B. G. Fry, M. Therese, B. Weckmueller *et al.*, "The family educational rights and privacy act of 1974," *Student records management: A handbook*, vol. 43, 1997.

[24] S. Ji, P. Mittal, and R. Beyah, "Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 1305–1326, 2017.

[25] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," in *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*. ACM, 2019, pp. 1–11.

[26] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

[27] J. A. Bartholomai and H. B. Frieboes, "Lung cancer survival prediction via machine learning regression, classification, and statistical techniques," in *2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2018, pp. 632–637.

[28] S. Walczak and V. Velanovich, "Improving prognosis and reducing decision regret for pancreatic cancer treatment using artificial neural networks," *Decision Support Systems*, vol. 106, pp. 110–118, 2018.

[29] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[30] H. Wimmer and L. Powell, "A comparison of the effects of k-anonymity on machine learning algorithms," in *Proceedings of the Conference for Information Systems Applied Research ISSN*, vol. 2167, 2014, p. 1508.

[31] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. ACM, 2016, pp. 308–318.

[32] F. D. Johansson, O. Frost, C. Retzner, and D. Dubhashi, "Classifying large graphs with differential privacy," in *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 2015, pp. 3–17.

[33] M. Nemati, J. Ansary, and N. Nemati, "Machine-learning approaches in covid-19 survival analysis and discharge-time likelihood prediction using clinical data," *Patterns*, vol. 1, no. 5, p. 100074, 2020.

[34] A. Spooner, E. Chen, A. Sowmya, P. Sachdev, N. A. Kochan, J. Trollor, and H. Brodaty, "A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.

[35] R. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," *Advances in neural information processing systems*, vol. 32, p. 9240, 2019.