DATA-DRIVEN PERSONALIZED CERVICAL CANCER RISK PREDICTION: A GRAPH-PERSPECTIVE

Vinay Chakravarthi Gogineni[⋆], *Severin R. E. Langberg*[†], *Valeriya Naumova*[⋆], *Jan F Nygård*[†], *Mari Nygård*[‡], *Markus Grasmair*[§], *and Stefan Werner*[‖]

[⋆]Dept. of Machine Intelligenc, SimulaMet-Simula Metropolitan for Digital Engineering, Norway
[†]Dept. of Registry Informatics, Cancer Registry of Norway, Norway
[‡]Dept. of Research, Cancer Registry of Norway, Norway
[§]Dept. of Mathematical Sciences, Norwegian University of Science and Technology, Norway
[‖]Dept. of Electronic Systems, Norwegian University of Science and Technology, Norway

ABSTRACT

Routine cervical cancer screening at regular periodic intervals leads to either over-screening or too infrequent screening of patients. For this purpose, personalized screening intervals are desirable that account for cancer risk development of individual patients. However, developing and training personalized risk prediction models is challenging since cancer screening data are scarce, irregular, and skewed. This paper proposes a personalized time-dependent cervical cancer risk prediction scheme using geometric deep learning (GDL) and spectral geometric matrix completion (SGMC) frameworks. The proposed approach learns row- and column-graphs from irregular and sparse cancer screening data. Then, we leverage the graph structure to reconstruct the continuous latent risk of individuals from screening data. During inference, the completed screening data matrix, comprising estimated individual continuous latent risk, serves as a dictionary for forecasting the cancer risk of new patients. We conducted experiments on synthetic and real-life screening data from the Cancer Registry of Norway.

1. INTRODUCTION

Various strains of the human papillomavirus (HPV) play a major role in causing cervical cancer, which develops cellular changes, from low-grade lesions to high-grade (pre-cancerous) lesions to invasive cancer [1]. Cervical cancer ranks third among the most common types of cancer for Norwegian women of ages 25 to 49. It is estimated that approximately 1% of Norwegian women will develop cervical cancer by the age of 75 [2]. The risk of developing cervical cancer can be observed from screening tests, e.g., cytology, histology, or HPV tests [3]. To this end, a class of mass-screening programs against cervical cancer have been established and successfully prevented 80% of cancer cases in the Nordic countries [4]. The current Norwegian Cervical Cancer Screening Program (NCCSP) recommends routine screening every third year from the age of 25 to 69 years, i.e., a total of 15 screenings if all would be normal. Since the risk of being infected with HPV and carcinogenesis varies significantly between females and also over time [5], the routine screening leads to a large number of unnecessary screening tests for patients unlikely to develop the disease (over-screening) and too infrequent screening of patients at high risk, resulting in under-treatment of pre-cancerous lesions [6]. Hence, our goal is to avoid the over-screening or under-treatment problem by predicting the personalized time-dependent risk of cervical cancer development.

The screening data collected through NCCSP contain the results of three clinical examinations: cytology, histology, and HPV, with no other medical information. These test results are categorized into four states that reflect the risk of receiving a future cancer diagnosis and the clinical consequences [7]. The first state, the *normal* state, indicates an accepted baseline risk. The second state, the *low-risk* state, indicates an early stage of carcinogenesis, which warrants more frequent screening to catch the disease before becoming invasive, but no treatment is usually required. The third state, the *high-risk* state, indicates a high probability for further cancer progress and requires immediate treatment. Finally, a *cancer* state is a failure of the screening program and a potential state for a patient. The large amount of screening data centrally organized at nationwide registries motivated us to use modern deep learning (DL) strategies for developing personalized preventive strategies that offer similar or increased protection at reduced over-screening and expenditure.

Sparse cervical screenings lead to a small fraction of the entries in the screening data matrix, whose rows represent screening history and columns represent time points in age. Furthermore, the recommendations are not strictly adhered to, and the individual screening histories are also irregular over time. In addition to the above, the screening data is also skewed, i.e., most screening results are normal. Specifically, the screening results in the NCCSP dataset are 94% normal, 4-5% low-risk, and around 1% are high-risk or cancer [8]. Personalized cancer risk prediction models have to be developed from this highly challenging data. To develop personalized screening models, we first reconstruct the underlying continuous latent risk of individuals from their partially observed screening histories using matrix completion. Then, the completed cancer screening data matrix is used as a dictionary to predict future screening outcomes of new patients to assist screening experts in recommending screening intervals.

Recently, geometric deep learning (GDL) [9, 10] and spectral geometric matrix completion (SGMC) [11] frameworks extensively pursued the matrix completion problem from a graph-perspective, where the graphs encode the relations among the rows or columns of the incomplete matrix. By imposing the smoothness priors on the row- and column-graphs, these above-mentioned techniques efficiently perform the task of matrix completion. However, for the cancer screening data, row- and column-graphs are not available. To

This work was supported by the Research Council of Norway.

this end, we learn a row-graph under smoothness constraints [12,13], i.e., screening histories are assumed to be smooth over the graph to be estimated. Furthermore, we generate a column-graph under the assumption that the risk of cancer development does not change rapidly within a year. Leveraging these geometric structures, we first reconstruct the continuous screening profiles from the partially observed screening histories through GDL and the SGMC frameworks. We then predict the future cancer development risk of an individual by computing and maximizing the conditional probabilities of the possible states. Finally, we apply the proposed approach on synthetic and real datasets to demonstrate its ability in predicting the future risk of cancer development.

2. PERSONALIZED CERVICAL CANCER RISK PREDICTION

In this work, the cervical cancer screening data is represented as a highly sparse matrix $\mathbf{X} \in \mathbb{N}^{N\times T}$, where each row $\mathbf{x}_n$ is the partially observed screening history of the nth patient Recall that the NCCSP currently recommends 3-year screening intervals for healthy patients and three to six months for patients at low risk. Hence, to provide a reasonable tradeoff between the temporal resolution and sparsity of the data, a three-month interval is considered for the time discretization of the data. In the following, Ω denotes the set of indices where $x_{n,t} \in \{1, 2, 3, 4\}$ is an integer encoding the observed states during the screening exam labeled normal, low-risk, high-risk of a cancer diagnosis, and actual cancer.

2.1. Reconstructing the Latent Risk of Cervical Cancer

The discrete observed states of a patient $X_{n,t}$ are assumed to be noisy measurements of an underlying continuous latent risk $Y_{n,t}$ that evolves slowly over time. In particular, we assume that an observed state is sampled from a discrete Gaussian distribution with mean $Y_{n,t}$ and variance $\frac{1}{2\theta}$. Here, the parameter $\theta > 0$ models the reliability of the risk assessment, which is given by

$$P(X_{n,t} = s \mid Y_{n,t}) = C_{Y_{n,t}} \exp(-\theta(s - Y_{n,t})^2), \qquad (1)$$

where $C_{M_{n,t}}$ is a risk-dependent normalization constant and s denotes the state of screening result. We estimate the underlying continuous latent risk from the partially observed screening histories by taking the recourse to the philosophy of geometric matrix completion. By constraining the space of solutions to be smooth w.r.t. certain geometric structure on the rows and columns of the screening matrix, the geometric matrix completion approach [14] can estimate the underlying continuous latent risk, i.e., $\hat{Y}_{n,t}$.

Let the undirected weighted row-graph $\mathcal{G}_r$ encode similarities among screening histories and the column-graph $\mathcal{G}_c$ encode the temporal smoothness of screening histories. Matrices $\mathbf{L}_r = \mathbf{\Phi}\mathbf{\Lambda}_r\mathbf{\Phi}^{\mathrm{T}}$ and $\mathbf{L}_c = \mathbf{\Psi}\mathbf{\Lambda}_c\mathbf{\Psi}^{\mathrm{T}}$ are the associated graph Laplacians. Note that the eigenbases of graph Laplacians represent the signals living on these graphs [15]. By exploiting the geometric structures and partially observed screening data matrix $\mathbf{X}$, the geometric matrix completion approach recovers the matrix $\mathbf{Y}$ that contains the continuous screening profiles of an individual, solving the following optimization problem [14]:

$$\min_{\mathbf{Y}} \|\mathbf{P}_\Omega \odot (\mathbf{Y} - \mathbf{X})\|_{\mathrm{F}}^2 + \frac{\gamma_r}{2}\|\mathbf{Y}\|_{\mathbf{L}_r}^2 + \frac{\gamma_c}{2}\|\mathbf{Y}\|_{\mathbf{L}_c}^2, \qquad (2)$$

where the symbol $\odot$ is the Hadamard product operator and $\mathbf{P}_\Omega : \mathbb{R}^{N\times T} \to \mathbb{R}^{N\times T}$ is a projection onto the masked entries, which sets all matrix entries not contained in Ω to 0 and the others unchanged. The quantity $\|\mathbf{Y}\|_{\mathbf{L}_r}^2 = \mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{L}_r\mathbf{Y})$ is the Dirichlet norm that quantifies the smoothness of screening profiles over the row graph. Similarly, the quantity $\|\mathbf{Y}\|_{\mathbf{L}_c}^2 = \mathrm{tr}(\mathbf{Y}^{\mathrm{T}}\mathbf{L}_c\mathbf{Y})$ quantifies the smoothness of each screening profile along time. Here, $\mathrm{tr}(\cdot)$ denotes the trace of its argument matrix. The regularization coefficients $\gamma_r, \gamma_c > 0$.

2.1.1. Separable Recurrent Multi-Graph Convolution Neural Network (sRMGCNN)

To reduce computational complexity and make it more suitable for large datasets, we adopt the factorized model for solving the optimization problem (2). Under the assumption that the latent risk profile of each patient is a linear combination of a small number of basic profiles $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ with $r << \min\{N, T\}$, the matrix $\mathbf{Y}$ of all such risks can be approximately decomposed as $\mathbf{Y} = \mathbf{U}\mathbf{V}^{\mathrm{T}}$ with $\mathbf{v} \in \mathbb{R}^{T\times r}$ is the collection of basic risk profiles and $\mathbf{U} \in \mathbb{R}^{N\times r}$ contains the patient-specific coefficients. By introducing the factorization model, the aforementioned geometric matrix completion problem in (2) reduces to [9]:

$$\min_{\mathbf{U},\mathbf{V}} \|\mathbf{P}_\Omega \odot (\mathbf{U}\mathbf{V}^{\mathrm{T}} - \mathbf{X}\|_{\mathrm{F}}^2 + \frac{\gamma_r}{2}\|\mathbf{U}\|_{\mathbf{L}_r}^2 + \frac{\gamma_c}{2}\|\mathbf{V}\|_{\mathbf{L}_c}^2. \qquad (3)$$

The factorized approach decouples the regularization that operates simultaneously in both rows and columns of $\mathbf{Y}$ in (2). To solve the optimization problem given in (3), we use geometric deep learning concepts. In particular, we use separable recurrent multi graph convolutional neural networks (sRMGCNN) formulated in the graph spectral domain [9]. The sRMGCNN solves the matrix completion problem in two stages. Firstly, one-dimensional multi-raph convolutions (MGC) will be applied to each factor w.r.t its graph. Let $T(\overline{\mathbf{L}}_r)$ and $T(\overline{\mathbf{L}}_c)$ be the respective Chebyshev polynomial of scaled Laplacians of the row- and column-graphs; then, the functionality of the multi-graph convolution neural network (MGCNN) layer can be described as [9]:

$$\widetilde{\mathbf{u}}_l = \sigma\Big(\sum_{l'=1}^{q}\sum_{j=0}^{p} \theta_{j,ll'}^r\ T(\overline{\mathbf{L}}_r)\ \mathbf{u}_{l'}\Big), \qquad (4)$$

and

$$\widetilde{\mathbf{v}}_l = \sigma\Big(\sum_{l'=1}^{q}\sum_{j=0}^{p} \theta_{j,ll'}^c\ T(\overline{\mathbf{L}}_c)\ \mathbf{v}_{l'}\Big), \qquad (5)$$

where $\theta_{j,ll'}^r$ and $\theta_{j,ll'}^c$ are the filtering coefficients in the MGCNN layer. In the next step, these extracted spatial features from MGCNN layers will be feeding to the recurrent neural network (RNN) that progressively reconstructs the complete screening profiles matrix by implementing a diffusion process. The sRMGCNN uses an LSTM architecture to learn complex non-linear diffusion processes [9]. The MGCNN, together with LSTM predict accurately small changes of the factors $\mathbf{U}$, $\mathbf{V}$ that can propagate through the full temporal steps.

2.1.2. Spectral Geometric Matrix Completion (SGMC)

In order to solve the problem in (2), the SGMC approach assumes the matrix $\mathbf{Y}$ is a permuted version of certain low-rank matrix $\mathbf{Z}$, i.e., $\mathbf{Y} = \mathbf{A}\mathbf{Z}\mathbf{B}^{\mathrm{T}}$ [11]. The SGMC interprets that the signal $\mathbf{Z}$ is sitting on a latent product graph factors$\mathcal{G}_r'$ and $\mathcal{G}_c'$ and it is transported onto the reference graph factors $\mathcal{G}_r$ and $\mathcal{G}_c$ via the linear transformation $\mathbf{A}\mathbf{Z}\mathbf{B}^{\mathrm{T}}$. By parameterizing the screening data matrix $\mathbf{Y}$ with

this matrix product, SGMC reduces the geometric matrix completion problem (2) to that of minimizing [11]:

$$\min_{\mathbf{A},\mathbf{Z},\mathbf{B}} \|\mathbf{P}_\Omega \odot (\mathbf{AZB}^{\mathrm{T}} - \mathbf{Y}\|_{\mathrm{F}}^2 + \frac{\gamma_r}{2}\|\mathbf{AZB}^{\mathrm{T}}\|_{\mathbf{L}_r}^2 + \frac{\gamma_c}{2}\|\mathbf{AZB}^{\mathrm{T}}\|_{\mathbf{L}_c}^2. \tag{6}$$

Note that the latent graph factors $\mathcal{G}_r'$ and $\mathcal{G}_c'$ illustrate the geometric interpretation, and there is no need to find them explicitly. Denoting the Laplacians of the latent graph factors by $\mathcal{L}_r'$, $\mathcal{L}_c'$ and their eigenbases by $\mathbf{\Phi'}$, $\mathbf{\Psi'}$, then, using these eigenbases and the eigenbases of reference Laplacians $\mathcal{L}_r$, $\mathcal{L}_c$, the factors can be expressed as [11]:

$$\mathbf{Z} = \mathbf{\Phi'C\Psi'}^{\mathrm{T}}, \mathbf{A} = \mathbf{\Phi P\Phi'}^{\mathrm{T}}, \text{and } \mathbf{B} = \mathbf{\Psi Q\Psi'}^{\mathrm{T}}. \tag{7}$$

Using (7), the screening data matrix $\mathbf{Y}$ can be reparameterized as $\mathbf{Y} = \mathbf{AZB}^T = \mathbf{\Phi PCQ}^{\mathrm{T}}\mathbf{\Psi}^{\mathrm{T}}$. Therefore, using this reparameterization of $\mathbf{Y}$, (2) becomes [11]:

$$\min_{\mathbf{P},\mathbf{C},\mathbf{Q}} \|\mathbf{P}_\Omega \odot (\mathbf{\Phi PCQ}^{\mathrm{T}}\mathbf{\Psi}^{\mathrm{T}} - \mathbf{Y}\|_{\mathrm{F}}^2 + \frac{\gamma_r}{2}\mathrm{tr}\big(\mathbf{QC}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{\Lambda}_r\mathbf{PCQ}^{\mathrm{T}}\big) + \frac{\gamma_c}{2}\mathrm{tr}\big(\mathbf{PCQ}^{\mathrm{T}}\mathbf{\Lambda}_c\mathbf{QC}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\big). \tag{8}$$

The above optimization problem can be solved using a gradient-descent.

2.1.3. Learning the Graphs

For reconstructing the continuous screening profiles, GDL and SGMC frameworks require row- and column-graphs. Since the graphs are not readily available with the screening data, we need to learn them from the partially observed screening data. Of these, we learn the row-graph under smoothness constraints, i.e., the underlying continuous latent risk is smooth over the row-graph. In other words, we assume that certain patients in the population exhibit similar screening histories. Under this assumption, the graph structure can be obtained by solving the following optimization problem [12, 13]:

$$\min_{\mathbf{W}_r \in \mathcal{W}_m} \sum_{t=1}^{T} \mathbf{x}_t^{\mathrm{T}}\mathbf{L}_r\mathbf{x}_t + f(\mathbf{W}_r), \tag{9}$$

where $\mathbf{x}_t$ is the tth column of the partially observed screening data matrix $\mathbf{X}$, $\mathbf{W}_r$ is the weighted adjacency matrix of the row-graph and $\mathcal{W}_m$ is the space that contains all valid non-negative, symmetric weighted adjacency matrices, i.e., $\mathcal{W}_m = \{\mathbf{W}_r \in \mathbb{R}_+^{N\times N} : \mathbf{W}_r = \mathbf{W}_r^{\mathrm{T}}, \mathrm{diag}(\mathbf{W}_r) = 0\}$. The regularization function $f(\mathbf{W}_r)$ in (9) prevents $\mathbf{W}_r$ being a zero matrix. Furthermore, the first term in (9) quantifies the smoothness of the screening histories over the selected graph, and it can also be expressed as $\sum_{t=1}^{T} \mathbf{x}_t^{\mathrm{T}}\mathbf{L}_r\mathbf{x}_t = \frac{1}{2}\mathrm{Tr}(\mathbf{W}_r\mathbf{R}) = \|\mathbf{W}_r \odot \mathbf{R}\|_1$, where $\mathbf{R}$ is the pairwise distance matrix. Researchers used various functions for $f(\mathbf{W}_r)$ in the literature. However, we follow [12] to obtain the sparse solution which is important in the case of large scale applications, and use $f(\mathbf{W}_r) = -\alpha\mathbf{1}^{\mathrm{T}}\log(\mathbf{W}_r\mathbf{1}) + \beta\|\mathbf{W}_r\|_{\mathrm{F}}^2$ with $\alpha \geq 0$ and $\beta \geq 0$. The logarithmic barrier forces the node degrees to be positive and the parameter β helps to control the sparsity of the graph, i.e., as β decreases, the solution of (9) becomes more sparse. The primal-dual techniques [16] can be used to obtain a solution for the optimization problem stated in (9). We construct the column-graph under the assumption that the risk of cancer development does not change rapidly within a year.

2.2. Predicting the Risk of Cervical Cancer Development

In this section, we present a method for predicting the future state of a patient given her screening record $\mathbf{x}$, i.e., the screening results from $t_1, \cdots, t_k$. We then predict the future state $x_{\hat{t}}$, for $\hat{t} > t_k$, by computing and maximizing the conditional probabilities of the possible states based on the model (1). For this purpose, the empirical distribution of reconstructed latent risk $\hat{Y}_{n,\hat{t}}$ is used in the place of their true distribution. Therefore, the conditional probabilities for the future state $x_{\hat{t}}$ are given by

$$p(x_{\hat{t}} = s \mid \mathbf{x}) \propto \sum_{n=1}^{N} C_{\hat{Y}_{n,\hat{t}}} \exp(-\theta(s - \hat{y}_{n,\hat{t}})^2) \times \prod_{j=1}^{k} C_{\hat{Y}_{n,t_j}} \exp(-\theta(y_{t_j} - \hat{y}_{n,t_j})^2). \tag{10}$$

Using (10), the conditions probabilities have to be calculated for $\forall s \in \{1, 2, 3, 4\}$. Then, the predicted state is the one with the maximum conditional probability.

3. NUMERICAL SIMULATIONS

This section demonstrates the ability of various algorithms for predicting the future risk of cervical cancer development. For this, numerical experiments were conducted on synthetic and real-life screening data. In all these experiments, the continuous latent risk matrix was reconstructed from training data and used it to estimate the conditional probabilities as described in Section II-B. Finally, these conditional probabilities have been used to predict the future risk at specific time points in independent test data. Each dataset was partitioned into 80% training and 20% test samples. The hyperparameters were optimized through cross-validation. In order to measure the quality of the prediction task, the k-category Matthews correlation coefficient (MCC) [17] was taken as a performance metric, which is given by

$$\mathrm{MCC}_k = \frac{m_+ m - \sum_{s=1}^{4} p_s t_s}{\sqrt{\big(m^2 - \sum_{s=1}^{4} p_s^2\big)\big(m^2 - \sum_{s=1}^{4} t_s^2\big)}} \tag{11}$$

In the above, m is the total number of test samples, and the number of correctly predicted test samples is m^+. Furthermore, the quantities t_s and p_s are the number of times a state s truly occurred and correctly predicted, respectively. The Matthews correlation coefficient summarizes the confusion matrix by a number $\mathrm{MCC}_k \in [-1, 1]$.

In the task of predicting the future risk of cancer development, we used forward fill (FF) (in which the last screening result is repeated to fill the missing screening result), sRMGCNN, and SGMC for matrix completion. The weighted row-graph utilized in the sRMGCNN and SGMC was obtained by solving (9). For comparative assessment, the same prediction exercise has also been carried out by matrix factorization (MF) [18] approach.

3.1. Results for Synthetic Data

The synthetic data was generated so that it resembles the high sparsity, randomness, and imbalance of the screening data. A latent risk matrix $\mathbf{Y} = \mathbf{UV}^{\mathrm{T}}$ is synthesized from a rank-five basis of the form $V_{t,k} = \exp(-10^3(t - \mu_k)^2)$ with $\mu_k \in \{70, 95, 120, 145, 170\}$ and

the patient-specific coefficients drawn from an exponential distribution. Combining inversion sampling with model (1) at $\theta = 2.5$, the real-valued risk matrix entries $Y_{n,t} \in [1, 4]$ are the integers representing the cancer risk states. This gives a complete state matrix $\mathbf{S}$, from which we derive a partially observed matrix $\mathbf{X} = \mathbf{P}_\Omega(\mathbf{S})$, by using inversion sampling with probabilities from the NCCSP data of observing an entry $S_{n,t}$ forward in the time given the result of the previous test to simulate Ω. For instance, a low-risk result calls for a follow-up screening within the next 3-6 months, whereas a high-risk result calls for immediate treatment. We further extend the data generation model by truncating each sample at times for the first and last screening, derived from the empirical distribution of the NCCSP data.

We repeat the procedure of synthesizing six datasets of each having 10000 samples with similar density (i.e., the fraction of observed entries in $\mathbf{X}$) for five different random seeds. We compared the models to an Oracle, which returns the most likely screening result given the true latent risk $Y_{n,t}$; with the error model (1); this amounts to rounding the true latent risk to the nearest integer. The MCC_k scores of various models vs. dataset density are illustrated in Fig. 1.

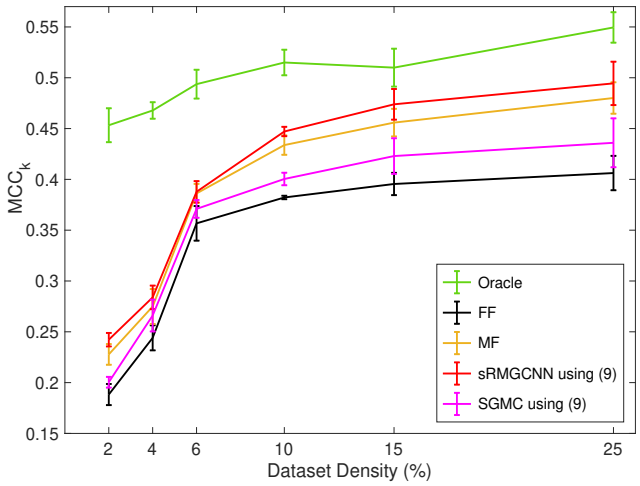

Fig. 1. Performance of various approaches on synthetic data given as the K-category Matthews correlation coefficient (MCC_k) against dataset density.

From Fig. 1, we see that the performance of all approaches is proportional to the synthetic dataset density. Furthermore, sRMGCNN based approach exhibited better performance on synthetic datasets compared to the FF, MF, and SGMC. The MGCNN, together with LSTM helped to predict the risk of cancer development, efficiently. Since the graphs learned from low-density datasets were poor in encoding the relations among screening histories, the graph-based approaches exhibited very poor performance on the low-density synthetic datasets.

3.2. Results for Screening Data

From the NCCSP population-level data, we randomly selected 10000 female patients with more than one exam. This dataset contains a median of 8 and at most 37 screening exams per patient; this yields a density of about 2.3% observed entries. Each history was aligned over a time grid spanning from youngest to oldest screened patient with a temporal resolution of three months. If multiple screenings were conducted within a three-month interval, we used the last observation to reflect the data available to clinicians when a prediction must be made. The MCC_k scores for predicting the future risk one to three years ahead in time from the NCCSP data are given in Table 1.

Table 1. Performance of various approaches given as the K-category Matthews correlation coefficient (MCC_k) on the NCCSP data

Forecast (years)	FF	MF	sRMGCNN using (9)	SGMC using (9)
1	0.1505	0.1250	0.1649	0.1646
2	0.0804	0.0728	0.1407	0.1473
3	0.0834	0.0429	0.1215	0.1458

From Table 1, it can be observed that the graph-based approaches perform slightly better than the FF and MF approaches. Furthermore, the graph-based approaches perform consistently over longer forecasts compared to FF and MF approaches. However, all these methods struggle with predicting the risk of cervical cancer development in a real-life scenario. The highly skewed data, i.e., very few cases with high-risk and cancer, is limiting the performance.

4. CONCLUSIONS

In this paper, we considered the problem of predicting the future risk of cervical cancer development in an individual. To this end, we have taken recourse to the geometric matrix completion concepts to reconstruct the continuous screening profiles of female patients from the partially observed screening histories. Then, the completed cancer screening data matrix has been used to forecast the cancer risk of a new patient. By leveraging the graph structure that encodes the similarities among the patients' screening histories, the matrix completion has been carried out through GDL and SGMC frameworks. Numerical experiments have been conducted both on synthetic and real-life screening datasets to demonstrate the potential of the proposed approach, and they revealed that the proposed approach could predict the short-term (12–36 months) individual risk of being diagnosed with cervical cancer, focusing on patients who would benefit from more frequent screening to reduce under-treatment.

5. ACKNOWLEDGEMENT

We would like to thank the authors of [11] for sharing their code.

6. REFERENCES

[1] P. A. Cohen, A. Jhingran, A. Oaknin, and L. Denny, "Cervical cancer," *Lancet*, vol. 393, no. 10167, pp. 169–182, 2019.

[2] Cancer Registry of Norway, "Cancer in Norway 2018 - Cancer incidence, mortality, survival and prevalence in Norway," *Cancer in Norway*, 2019.

[3] WHO. Cervical cancer. [Online]. Available: https://www.who.int/health- topics/cervical-cancer

[4] S. Vaccarella, S. Franceschi, G. Engholm, S. Lönnberg, S. Khan, and F. Bray, "50 years of screening in the Nordic countries: Quantifying the effects on cervical cancer incidence," *Br. J. Cancer*, vol. 111, no. 5, pp. 965-969, Aug. 2014.

[5] K. Pedersen, S. Fogelberg, L. H. Thamsborg, M. Clements, M. Nygard, I. S. Kristiansen, E. Lynge, P. Sparen, J. J. Kim, and E. A. Burger, "An overview of cervical cancer epidemiology and prevention in Scandinavia," *Acta obstetricia et gynecologica Scandinavica*, vol. 97, no. 7, pp. 795–807, 2018.

[6] K. Pedersen, E. A. Burger, S. Campbell, M. Nygard, E. Aas, and S. Lonnberg, "Advancing the evaluation of cervical cancer screening: Development and application of a longitudinal adherence metric," *The European J. Public Health*, vol. 27, no. 6, pp. 1089–1094, 2017.

[7] B. C. Soper, M. Nygard, G. Abdulla, R. Meng, and J. F. Nygard, "A hidden Markov model for population-level cervical cancer screening data," *Statistics in Medicine*, vol. 39, no.25, pp. 3569– 3590, 2020.

[8] J. F. Nygard, S. O. Thoresen, and G. B. Skare, "The cervical cancer screening program in Norway, 1992-2000: Changes in pap-smear coverage and cervical cancer incidence," *Int. J. Cancer*, pp. 110–110, 2002.

[9] F. Monti, M. Bronstein, and X. Bresson, "Geometric matrix completion with recurrent multi-graph neural networks," in *Proc. Advances in Neural Info. Process. Systems*, 2017, pp. 3697–3707.

[10] F. Monti, D. Boscaini, J. Masci, E. Rodolà, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model CNNs," in *Proc. Computer Vision and Pattern Recognition*, 2017.

[11] A. Boyarski, S. Vedula and A. Bronstein, "Deep matrix factorization with spectral geometric regularization," 2020. [Online]. Available: https://arxiv.org/abs/1911.07255

[12] V. Kalofolias, "How to learn a graph from smooth signals," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2016, pp. 920–929.

[13] X. Dong, D. Thanou, P. Frossard and P. Vandergheynst, "Learning Laplacian matrix in smooth graph signal representations," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6160-6173, Dec. 2016.

[14] V. Kalofolias, X. Bresson, M. M. Bronstein, and P. Vandergheynst, "Matrix completion on graphs," 2014. [Online]. Available: https://arxiv.org/abs/1408.1717

[15] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, 2013.

[16] N. Komodakis and J. Pesquet, "Playing with duality: An overview of recent primal-dual approaches for solving largescale optimization problems," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 31–54, 2015.

[17] J. Gorodkin, "Comparing two k-category assignments by a k-category correlation coefficient," *Computational Biology and Chemistry*, vol. 28, no. 5-6, pp. 367–374, 2004.

[18] J. Zhou, F. Wang, J. Hu, and J. Ye, "From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records," *Proc. Int. Conf. on Knowledge discovery and data mining*, 2014, pp. 135–144.