Computers in Biology and Medicine 149 (2022) 106048

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

GCRS: A hybrid graph convolutional network for risk stratification in multiple myeloma cancer patients

Dikshant Sagar [a,1], Priya Aggarwal [b,1], Akanksha Farswan [a], Ritu Gupta [c,*], Anubha Gupta [a,*]

[a] SBILab, Department of ECE, Indraprastha Institute of Information Technology, New Delhi, India
[b] Vehant Technology Pvt. Ltd., Noida, India
[c] Laboratory Oncology Unit, Dr. B.R.A. IRCH, AIIMS, New Delhi, India

ARTICLE INFO

Keywords:
Survival analysis
Graph convolutional neural networks
Risk stratification
Multiple myeloma
Artificial intelligence

ABSTRACT

In this study, we present an efficient Graph Convolutional Network based Risk Stratification system (GCRS) for cancer risk-stage prediction of newly diagnosed multiple myeloma (NDMM) patients. GCRS is a hybrid graph convolutional network consisting of a fusion of multiple connectivity graphs that are used to learn the latent representation of topological structures among patients. This proposed risk stratification system integrates these connectivity graphs prepared from the clinical and laboratory characteristics of NDMM cancer patients for partitioning them into three cancer risk groups: low, intermediate, and high. Extensive experiments demonstrate that GCRS outperforms the existing state-of-the-art methods in terms of C-index and hazard ratio on two publicly available datasets of NDMM patients. We have statistically validated our results using the Cox Proportional-Hazards model, Kaplan–Meier analysis, and log-rank test on progression-free survival (PFS) and overall survival (OS). We have also evaluated the contribution of various clinical parameters as utilized by the GCRS risk stratification system using the SHapley Additive exPlanations (SHAP) analysis, an interpretability algorithm for validating AI methods. Our study reveals the utility of the deep learning approach in building a robust system for cancer risk stage prediction.

1. Introduction

Multiple Myeloma (MM) is a hematological malignancy causing an unrestricted proliferation of clonal plasma cells in the bone marrow. The overall survival period of MM patients varies from 6 months to more than ten years, depending upon the risk stage of a patient at the time of diagnosis and the subsequent treatment given to the patient. The varying outcome and risk stages predicted in patients are mainly caused due to the clinical and biological heterogeneity in multiple myeloma. Recent advances in cancer genomics have facilitated the identification of tumor heterogeneity in MM patients [1], thereby optimizing a patient's treatment by examining their response to the treatment. In addition, risk stratification tools are increasingly used in oncology practice to identify high-risk patients who may require higher recurrent therapy visits to hospitals to prolong their overall survival. Multiple prognostic systems [2–6] have been described for MM. The first work in this direction was proposed by Salmon and Durie in 1975 [2] that grouped patients into three risk categories based on differential overall survival. Next, International Staging System (ISS) [3] was proposed by Greipp PR and colleagues in 2005, which utilized serum albumin and β2-microglobulin as parameters to identify the initial risk stage. However, with advances in treatment therapies, the landscape of survival changed drastically [7,8]. Further, the development of new techniques in biology allowed scientists to examine the genomic changes at molecular levels in MM, which facilitated the addition of chromosomal alterations in the staging system leading to the Revised-ISS (R-ISS) [4]. R-ISS utilizes serum albumin, β2-microglobulin, lactate dehydrogenase (LDH), and high-risk cytogenetic aberrations (HRCA) to predict the cancer risk stage. Nowadays, triplet combination therapy has become the new standard of care in MM, due to which many high-risk patients are now shifted to a low/intermediate-risk category. However, better care and treatment require the correct identification of the risk group of a patient.

Machine learning (ML) techniques have greatly benefited the biomedical domain and provide an excellent framework for efficiently utilizing and analyzing the abundance of medical/biomedical records to extract valuable information. Various machine learning methods have been devised for risk stratification in disorders such as in type-2 diabetic patients [9], in cardiovascular disorders [10], in prostate

[*] Corresponding authors.
E-mail addresses: drritugupta@gmail.com (R. Gupta), anubha@iiitd.ac.in (A. Gupta).
[1] Equal contribution.

https://doi.org/10.1016/j.compbiomed.2022.106048
Received 11 March 2022; Received in revised form 31 July 2022; Accepted 20 August 2022
Available online 24 August 2022

cancer [11], in resected gastric cancer [12] and in nasopharyngeal carcinoma [13]. Motivated by the relevance and versatility of ML/AI methods, two AI-enabled risk staging models: Modified Risk Staging (MRS) [14] and Consensus-based Risk-Stratification System (CRSS) [15], have recently been proposed for MM risk stratification. MRS is a robust AI-enabled method that uses six easy-to-acquire laboratory parameters: albumin, beta-2 microglobulin (β2m), calcium, Glomerular Filtration Rate (eGFR), hemoglobin, and age to predict the risk stages. MRS method has been observed to be superior to ISS and R-ISS in the predictions of progression-free survival (PFS) and overall survival (OS) [14].

Many studies have reported variations in the overall survival of groups across different race/ethnicity [16–19], thereby justifying the need for a race/ethnicity-specific risk-stratification system. Although ethnicity plays a vital role in predicting the risk for MM [20] and can significantly impact the risk score prediction, none of the earlier MM risk-staging systems (ISS, R-ISS, or MRS) integrated ethnicity-specific information. The CRSS method devised different parameter cutoffs based on the ethnicity information on the American and Indian cohorts [15]. It incorporated ethnicity-specific cutoffs of the above six laboratory parameters and considered an additional High-Risk Cytogenetic Aberrations (HRCA) parameter for risk staging. As a result, CRSS performed superior to all previous MM risk-staging systems, including ISS, R-ISS, and MRS.

Hence, we also incorporated ethnicity information using the same parameter cutoffs on normal and abnormal parameter ranges using those identified by the CRSS method on the American and Indian cohorts of NDMM patients.

It has been shown that correlations, similarities, and interactions exist between patients with similar characteristics or laboratory parameters. Keeping this in mind, the collective prognostic impact of the parameters was integrated into risk staging in the CRSS method via the creation of three different patients-to-patients adjacency graphs [15]. This helped in extracting a patient topology based on the assumption that patients with similar medical records would be closer to or connected in a graph structure and have a similar prognosis [15]. Later, ML methods of Gaussian Mixture Modeling (GMM) and agglomerate clustering were utilized to stratify the patients into three risk groups.

This study shows that this patient-to-patient adjacency graph network can be leveraged using a new generation deep learning method that inherently thrives on graphical data known as the Graph Convolutional Network (GCN). GCNs, in its essence, look at both the patient and its neighbors when making a decision using the graph Laplacian. We also propose a hybrid approach while training the network that uses a fusion of multiple adjacency graphs formed in different ways. We name the proposed method Graph Convolutional based Risk-Stratification System (GCRS). We also tackle the problem of an imbalanced dataset (i.e., having an unequal number of patients in three risk groups) while training the GCN to alleviate the problem of overfitting to one class. We also interpret results of GCRS using SHapley Additive exPlanations (SHAP) [21] analysis. There is no evidence of how accurately the features are extracted/learned by the deep learning models because they are generally treated as black-box classifiers. SHAP analysis helped us visualize important values assigned to input features that lead to the predicted outcome. We evaluated the performance of the proposed GCRS model on two datasets: the popular Multiple Myeloma Research Foundation (MMRF) dataset and another from the Institute Rotary Cancer Centre at All India Institute of Medical Sciences (AIIMS) introduced in [15], named as MMIn dataset. Our model achieves robust performance on both these datasets compared to the existing methods, including the recent MRS and CRSS methods. The salient contributions of this study are summarized below:

1. The patient-to-patient interrelationships have been leveraged using a graph-topology structure in the newly diagnosed MM patients to learn the inherent similarities between the patients belonging to the same risk stage and the differences between the patients belonging to different risk stages.
2. A hybrid graph view of the patient-to-patient adjacency graph network is employed by using three adjacency graphs generated in different ways.
3. A deep learning architecture exploits the above-generated graph structure using recent deep learning structures of GCNs for an improved and generalizable risk staging in newly diagnosed MM cancer patients.
4. The DL model's learning and inference are interpreted using the SHAP algorithm to understand the contributions of each parameter to the risk stages.
5. Risk stage validation of the proposed GCRS method is performed using the statistical analysis on the parameters of all the three predicted risk groups.
6. A thorough comparison of GCRS with the state-of-the-art methods is carried out using the Kaplan–Meier Curves and the Cox Proportional Hazard Ratios.

2. Materials and methods

2.1. Datasets

In this work, we have utilized two MM cancer datasets (Refer to Table 1). The first dataset is the MMIn dataset and consists of a total of 1070 Indian MM patients [14,15]. The Institute Rotary Cancer Centre has provided this data to the All India Institute of Medical Sciences (AIIMS). Further details about the selection of patients in this dataset are available in [14]. The second dataset has a total of 900 MM patients enrolled in the Multiple Myeloma Research Foundation (MMRF) project and is available publicly at the MMRF research gateway (https://research.themmrf.org/). Patients in this dataset belong to the American population. We utilized the data of 384 MMIn and 800 MMRF patients whose high-risk cytogenetic information was available. From both the dataset, seven features were used for the risk staging similar to [15], namely β2m, Age, Albumin, Calcium, Hemoglobin, eGFR, and high-risk cytogenetic abnormalities (a binary feature indicating the presence of t(4;14), t(14;16) and/or del17). The clinical significance of these features is as follows. Serum albumin and hemoglobin reflect the residual normal functioning of the hematopoietic system compromised by infiltration of malignant plasma cells; β2m reflects the tumor burden; calcium and creatinine reflect the bone and kidney homeostasis; and t(4;14), t(14;16) and/or del17 are known biomarkers of poor clinical outcome in MM. Since age also impacts the performance status and clinical outcome, it has been included in the risk stratification modeling in the present study. A total of 41 patients (3.8% of 1070) had one or two missing values in the MMIn dataset. Although multiple methods have been proposed in the literature to recover missing values [22,23], we imputed the missing values with the median value of the parameters because MMIn dataset had only a few missing values. Response outcomes were estimated in the MMIn dataset following the International uniform response criteria for multiple myeloma. Progression-free survival (PFS) and overall survival (OS) differ in terms of time computation from the date of diagnosis. In PFS, duration from the date of diagnosis till the date of progression is considered, whereas in OS, date of diagnosis till the date of death or the date of being censored is considered.

Generally, there are established thresholds (irrespective of ethnicity or race) for each of the above six parameters that are used to divide a patient into high-risk or low-risk groups. However, recently, these thresholds have been computed separately for both the MMRF and MMIn datasets using the K-adaptive partitioning algorithm [15]. For MMRF dataset, the calculated thresholds were 67 years for age, 9.59 g/dL for hemoglobin, 5.5 mg/L for β2m, 3.5 g/dL for albumin, 10.52 mg/dL for calcium and 48.3 mL/min/1.73 m^2 for eGFR. Similarly, for the MMIn dataset, thresholds were 67 years for age, 12.3 g/dL for hemoglobin, 4.78 mg/L for β2m, 3.5 g/dL for albumin, 11 mg/dL for calcium and 48.2 mL/min/1.73 m^2 for eGFR. These thresholds differed

Table 1
Total number of samples and features present in the MMRF and MMIn dataset.

	MMRF	MMIn
Total no. of patients	900	1070
No. of patients whose high risk cytogenetic feature is available	800	384
Features used for training	Age, Hemoglobin, β2m, Albumin, Calcium, eGFR and High Risk Cytogenetic Feature	Age, Hemoglobin, β2m, Albumin, Calcium, eGFR and High Risk Cytogenetic Feature

by ethnicity [15]. Compared to the established thresholds, the above-computed thresholds demonstrated better separability between the low- and the high-risk groups when assessed using the statistical log-rank test on the Kaplan–Meier curves in the earlier risk staging [15]. Thus, we have also utilized these new thresholds for different parameters in this work. Accordingly, the description of the two datasets is presented in detail in Table 2.

2.2. Proposed Graph Convolutional based Risk-Stratification system (GCRS)

This section explains the proposed GCRS risk-stratification system's complete workflow. The workflow consists of three steps, including (i) the computation of multiple adjacency graphs, (ii) assigning risk label/stage to each patient using the adjacency graphs, and finally, (iii) training a GCN using the patient-specific laboratory parameters and the assigned risk labels so that the trained network can be used to predict the risk stage of a new patient using the laboratory parameters. All three steps are discussed in detail below.

Step 1: Computation of multiple adjacency graphs
In this step, the ethnicity-specific thresholds computed earlier on the parameters of the MMIn and MMRF dataset are used to create three adjacency graphs [15]. For the sake of completeness, a summary is presented below.

(i) The first adjacency graph is created using Hazard Ratio (HR) values. These values were obtained from a univariate Cox hazard analysis. We obtained two HR values for each parameter, one from PFS and another from OS. The highest of the two HR values was considered, followed by normalization using min–max scaling with the scale ranging from 1 to 4. Weights to each of the parameters were assigned based on these scaled HR values, representing the impact of each parameter on patients' survival. This resulted in assigning a risk score to each patient. These scores were then used to compute an adjacency graph of size $n \times n$, where n denotes the number of patients. This adjacency graph is computed by taking absolute differences between the score of patients.

(ii) For the second adjacency graph, ranks obtained from the multivariate Cox hazard analysis were used. The highest and lowest ranks to the parameters were assigned based on the highest and lowest hazard values, respectively. Weights of each of the parameters were assigned based on these rank values, representing the impact of each parameter on a patient's survival. Further, the risk score for each patient was calculated by the successive addition of the weights of all parameters having values greater than the cutoffs defined for the high-risk group. These scores were then used to compute an adjacency graph of size $n \times n$ by computing absolute differences as mentioned above.

(iii) The third adjacency graph was created using the p-values obtained by performing a log-rank test on the Kaplan–Meier (KM) curves. We obtained two p-values for each parameter, one from PFS and another from OS. Lowest of the two p-values was considered, followed by normalization using min–max scaling with the scale range from 1 to 4. Weights of each of the parameters were assigned based on this scaled p-values. The risk score for each patient was calculated by the successive addition of the weights of all parameters having values greater than the cutoffs defined for the high-risk group. These scores were then used to compute an adjacency graph of size $n \times n$ using absolute differences.

Step 2: Assigning ground-truth risk labels to each patient
This step is required due to the absence of actual ground truth risk labels (low, intermediate, and high) in the collected medical data. Since, eventually, a risk stage predictor or a classifier needs to be built up that accepts the laboratory and clinical parameters and predicts the risk stage, ground-truth risk-stage labels of patients are required to train such a classifier. Similar to [15], GMM-based clustering was utilized to obtain clustering labels on all three adjacency graphs, followed by the creation of a consensus graph of the same size as that of the original adjacency graph. An (i, j)th entry in this graph at the ith row and jth column position was determined by calculating the number of times, the ith and the jth patients were assigned the same group. Diagonal entries were zero in this graph. Agglomerative clustering was subsequently performed on the consensus graph to cluster the patients into three risk groups. This allowed for the grouping of existing patients into three different risk stages as low, intermediate, and high based on the qualitative analysis of the medical data. The three risk groups are labeled as GCRS-1 (low-risk), GCRS-2 (intermediate risk), and GCRS-3 (high-risk). This grouping of existing patients was used as the ground truth to train a novel and scalable supervised algorithm for solving the problem of risk staging in MM and inferring risk labels on prospective patients once deployed in a hospital.

Step 3: Training of GCRS
Once all the three adjacency graphs were available along with the ground truth risk label of each patient, we trained a GCN to solve the three-class classification problem. We first summarize GCN, followed by the details of the proposed GCRS approach. We use bold uppercase letters for matrices, bold lowercase letters for vectors, and lowercase letters for scalars. We denote an entry of the ith row and jth column of matrix **A** by A_{ij}.

(a) *Brief on GCN:* Here, we present the notations and some preliminaries for the GCNs. Due to abundant graph-like data in real-life scenarios, research from the past few years has paved the way for extracting information from the graph data [24–29]. A GCN [30] is a convolutional neural network that operates directly on a graph, wherein an adjacency matrix **A** is used as a graph. This graph is a collection of nodes (patients in our case) and edges (connections between patients). Graph convolution operation assimilates the neighborhood features of adjacent nodes via utilizing the graph structure, which is similar to the vanilla convolution operation.

In GCN, every node is assumed to be connected to itself and is also connected with the other nodes based on the properties of their neighborhoods. Let $\mathbf{X} \in \mathbb{R}^{n \times m}$ be a feature matrix containing all the n nodes with their m features. Each row of this matrix is a feature vector of a particular node. The diagonal elements of matrix **A** are set to 1 because of self-loops. This can be achieved by adding the identity matrix **I** to **A** as $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. For a one-layer GCN, the new k-dimensional node feature matrix $\mathbf{H} \in \mathbb{R}^{n \times k}$ is computed as [See Fig. 1]:

$$\mathbf{H} = \sigma(\mathbf{D}^{-\frac{1}{2}} \tilde{\mathbf{A}} \mathbf{D}^{-\frac{1}{2}} \mathbf{X} \mathbf{W}), \tag{1}$$

Table 2
Baseline demographic, laboratory and clinical characteristics of multiple myeloma (MM) patients of the MMRF and MMIn cohort.

Parameters Cutoffs	MMRF (n = 900)	Parameters Cutoffs	MMIn (n = 1070)
Age	700 (77.77%)	Age	936 (87.47%)
≤69	200 (22.22%)	≤67	134 (12.53%)
>69		>67	
Hemoglobin (g/dL)	253 (28.11%)	Hemoglobin (g/dL)	912 (85.23%)
≤9.59	647 (71.88%)	≤12.3	158 (14.76%)
>9.59		>12.3	
Serum albumin (g/dL)	375 (41.66%)	Serum albumin (g/dL)	496 (46.35%)
≤3.5	525 (58.33%)	≤3.5	574 (53.64%)
>3.5		>3.5	
Beta 2 microglobulin (mg/L)	661 (73.44%)	Beta 2 microglobulin (mg/L)	457 (42.71%)
<5.5	239 (26.55%)	<4.78	613 (57.28%)
≥5.5		≥4.78	
Serum calcium (mg/dL)	784 (87.11%)	Serum calcium (mg/dL)	926 (86.54%)
<10.52	116 (12.88%)	<11	144 (13.45%)
≥10.52		≥11	
eGFR (ml/min/1.73 m^2)	163 (18.11%)	eGFR (ml/min/1.73 m^2)	347(32.42%)
≤48.3	737 (81.88%)	≤48.2	723(67.57%)
>48.3		>48.2	
ISS 1/2/3	342/319/239	ISS 1/2/3	207/323/540
R-ISS 1/2/3	107/505/91	R-ISS 1/2/3	47/459/121
MRS 1/2/3	337/408/155	MRS 1/2/3	281/511/278
CRSS 1/2/3	174/452/174	CRSS 1/2/3	137/192/55

Note: The HRCA parameter is not available for all the patients. ISS and MRS do not use the HRCA parameter and hence, their risk-stage labels are available for a larger set of patients.

where $\sigma(.)$ is the nonlinear activation function, $\mathbf{W} \in \mathbb{R}^{m\times k}$ is a weight matrix of trainable filter parameters, and $\mathbf{D}$ is the degree matrix corresponding to the adjacency matrix $\mathbf{A}$ which is computed as $D_{ii} = \Sigma_j A_{ij}$. One can incorporate the higher order neighborhood information by stacking multiple GCN layers as:

$$\mathbf{H}^{l+1} = \sigma(\mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\mathbf{H}^{l}\mathbf{W}^{l}), \tag{2}$$

where l denotes the layer number and $\mathbf{H}^0 = \mathbf{X}$.

(b) *Proposed GCRS model training:* As mentioned in the GCN details above, it requires one adjacency graph and the feature matrix as an input. Therefore, we fused all the three adjacency graphs to form a single adjacency graph using a α-weighted sum fusion given by:

$$\mathbf{A} = \alpha\mathbf{A}_1 + \alpha^2\mathbf{A}_2 + \alpha^3\mathbf{A}_3, \tag{3}$$

where an optimal $\alpha = 0.7$ was found through a grid search method while training the network. The resulting fused adjacency graph is passed through the proposed GCRS model (Fig. 2). Our proposed GCRS model consists of two input layers that accept the features $\mathbf{X}$ and the adjacency graph $\mathbf{A}$. These are fed to the GCN Layer with $F = 16$ hidden units and a Rectified Linear Unit (ReLU) as an activation function. GCN maps the convolved and pooled features to a dimension of $n\times 16$, which is further fed to the last (output) layer for prediction via a softmax activation function (See Fig. 2). So the model equation becomes:

$$\mathbf{Z} = Softmax\Big(ReLU(\mathbf{D}^{-\frac{1}{2}}\tilde{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}\mathbf{X}\mathbf{W})\Big), \tag{4}$$

where $Softmax(x_c) = \frac{exp(x_c)}{\sum_{j=1}^{C} exp(x_j)}$ with $c = 1, \dots, C$ and C is the number of classes. The binary cross-entropy loss function is used as the objective function to be minimized over all the samples and is defined as:

$$L = -\sum_{i=1}^{N}\left[\sum_{c=1}^{C} y_i(c)\, log\Big(p_i(c)\Big)\right], \tag{5}$$

where N is the number of training samples, C is the number of classes (which is equal to 3 in our case), $y_i(c)$ is the binary indicator that is ‘1’ if the class label c is the true label of observation i and is ‘0’ otherwise, and $p_i(c)$ is the predicted probability with which the observation i belongs to class c.

We observed an unequal number of patients in ground-truth risk groups. This could lead to overfitting of the model to the majority class and hence, could result in a biased model. In order to handle the class imbalance problem present in both MMRF and MMIn datasets, we employed a weighted loss calculation that assigns a higher weight to the loss of minority classes and a lower weight to the majority class. The weight w_c for class c is calculated as:

$$w_c = \frac{N}{C \times N_c}, \tag{6}$$

where N is the number of training samples, C is the number of classes and N_c is the number of samples belonging to the class c.

2.3. Drawing inferences on a prospective subject using the GCRS model

At the time of inference (or testing/using the trained model) on a new patient, the adjacency matrix is again constructed by adding one additional row and column to the existing adjacency (constructed on the training dataset), which signifies the new patient’s connections in the current patient graph via the three methods explained in Step 1. To have consistent dimensions for the model to work, the new patient’s feature vector is stacked with the training dataset in $\mathbf{X}$ and forward passed with the modified adjacency matrix through the GCN layer using (4) of the GCRN model (Refer to Fig. 2) to obtain the class or the label of the new patient.

3. Results

The proposed GCRS model was trained on Google Colab. The model based on the MMRF dataset (800 patients having the HRCA information) has been trained for a maximum of 3000 epochs as it achieved a loss saturation at ∼3000 epochs (See Fig. 3(a)). The trained model is then fine-tuned on the MMIn dataset (384 patients) using transfer learning (TL) of the trained MMRF model weights. We used transfer learning because pre-trained weights have been known to provide

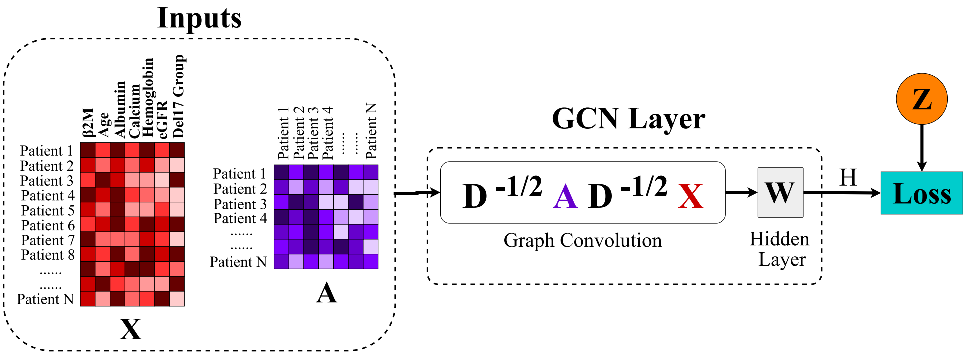


Fig. 1. A single GCN layer schema, where **X** is an input feature matrix, **A** is an adjacency matrix, **D** is the degree matrix calculated from **A**, **W** is a weight matrix, **H** is the layer's output and **Z** is the ground truth for loss calculation.

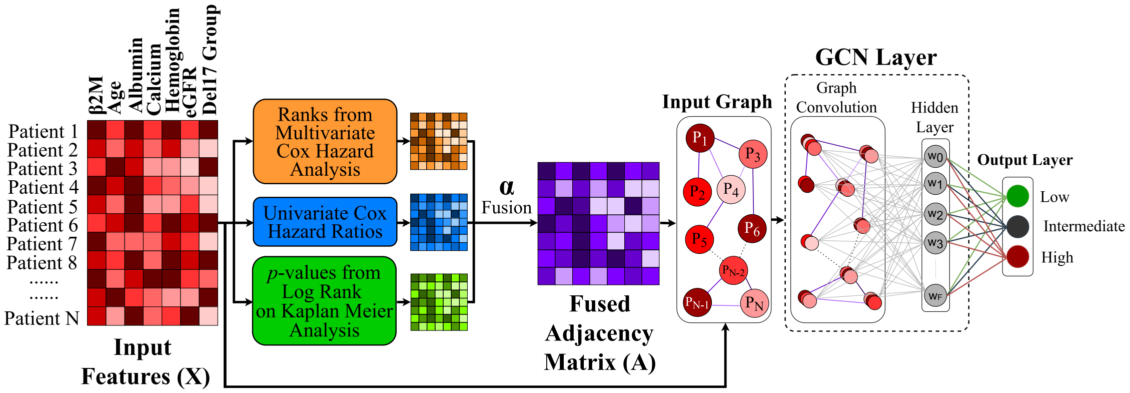


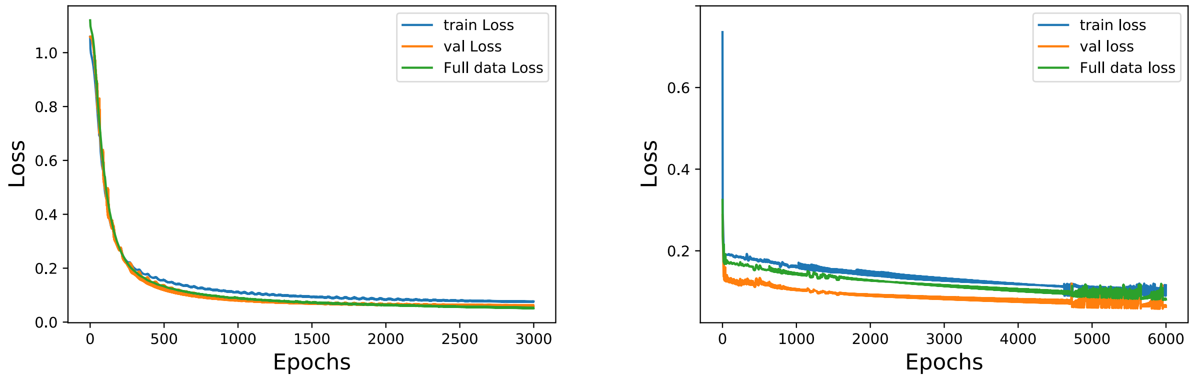
Fig. 2. GCRS Overview: Our proposed GCRS model consists of two input layers that accept features **X** and the adjacency graph **A**. These are fed to the GCN Layer with hidden units and a Rectified Linear Unit (ReLU) as an activation function. GCN maps the convolved and pooled features and feeds it to the last (output) layer for prediction via a softmax activation function. The last layer predicts the risk stage of the patient as the final output.

much-stabilized training and faster convergence for the same objective function [31]. With TL, the loss saturated at ∼6000 epochs (See Fig. 3(b)). Since we have a small number of data samples in the MMIn dataset and it is generally known that a low number of training samples in a supervised neural network training can result in a poor approximation and a high variance when trained for a fewer number of epochs [31], we observe that a relatively higher number of epochs were needed to train our network.

To ensure the absence of overfitting, both datasets were split with a ratio of 80:20 to create train and validation sets. Their respective loss plots while training GCRS on the train set can be seen in Fig. 3. Since the datasets are small, the models were trained on the whole datasets for the best model possible. This is to note that the loss plot of 100% data also shows a similar curve as the 80–20 split as seen in Fig. 3. This figure indicates the absence of overfitting during training and shows that the models are correctly trained on both datasets. Moreover, this emphasizes that the inference strategy on the GCNs differs from the traditional DL models. As described in [32], a new adjacency is calculated for the total data (together of the training dataset and the test dataset) at the time of inference. Thus, again the entire data is utilized at the time of inference. This adjacency matrix is passed through the model to obtain the labels of the test set. We used the labels obtained on the dataset to calculate the KM Curves and C-Indexes (Hazard Ratios).

The hyperparameters used to train the model are as follows: Adam (Optimizer), 0.1 (β_1), and 0.01 (Learning Rate). The proposed GCRS method requires initialization of α parameter as defined in (3) for the fusion of the adjacency matrices and the training parameters such as the learning rate for the optimizer and the number of epochs for training the graph convolutional neural network. Optimal values of all these parameters were found using grid search with validation set classification loss as an objective to obtain the best hyperparameters. This is one of the standard approaches for obtaining the values of hyperparameters in ML/DL methods. In this search approach, all possible combination values of the hyperparameters are run through so as to obtain a final objective function value. Furthermore, the values of the hyperparameters on which the objective function has the minimum value are returned as optimal. In this paper, the objective function for grid search was the minimization of the proposed GCRS model's classification loss. Packages called keras tuner and sklearn were used to run this grid search algorithm to find the best hyperparameter values. Fig. 4(b) shows the plots of training loss versus epoch with $\alpha = 0.7$ that yielded minimum loss and hence, better risk staging performance.

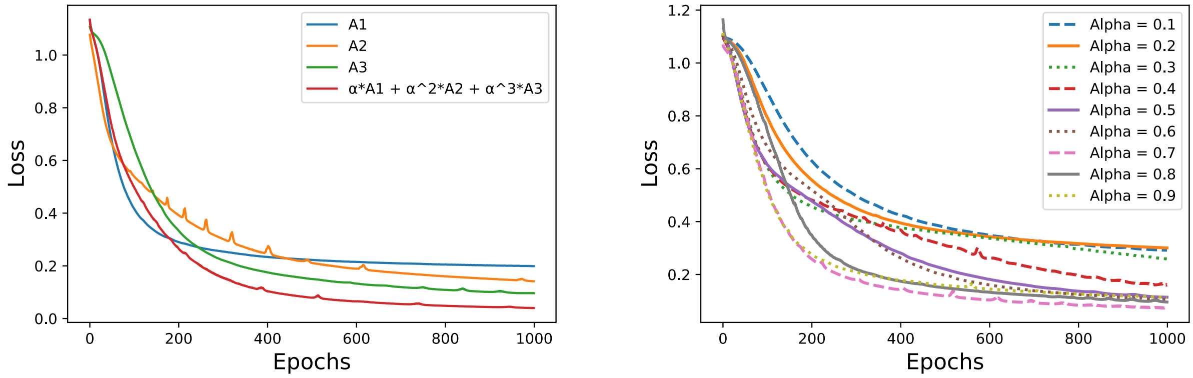
It was observed that the largest proportion of patients in both MMRF ($n = 800$) and MMIn ($n = 384$) datasets got assigned to GCRS-2 (MMRF: $n = 439$, 54.86%, MMIn: $n = 195$, 50.78%). This was followed by GCRS-3 ($n = 187$, 23.38%) and then GCRS-1 ($n = 174$, 21.75%) in the MMRF dataset. Similarly, the numbers of the GCRS-2 group were followed by GCRS-1 ($n = 132$, 34.38%) and then GCRS-3 ($n = 57$, 14.84%) in the MMIn dataset. Performance analysis of the proposed GCRS model is carried out using the confusion matrix. Confusion matrices of both MMRF and MMIn datasets are shown in Fig. 5. The diagonal elements of these matrices show the number of correct classifications in each category. The performance concerning each class can be understood with the help of the confusion matrix. It is observed that the DL model correctly identifies the low-risk group of patients. However, the trained model changed some risk labels of the intermediate and high-risk patients. The DL model might have learned better separability among the different risk groups based on their laboratory parameters because it is a richer model than the clustering carried out in Step 2 of Section 2.2. In the next subsection, we perform the statistical analysis and analyze the Kaplan–Meier curves on these identified risk labels to assess the DL model's performance in identifying patients' risk labels.



(a) Loss curves of GCRS on the MMRF dataset for the trained model.

(b) Loss curves of GCRS on the MMIN dataset for the trained model.

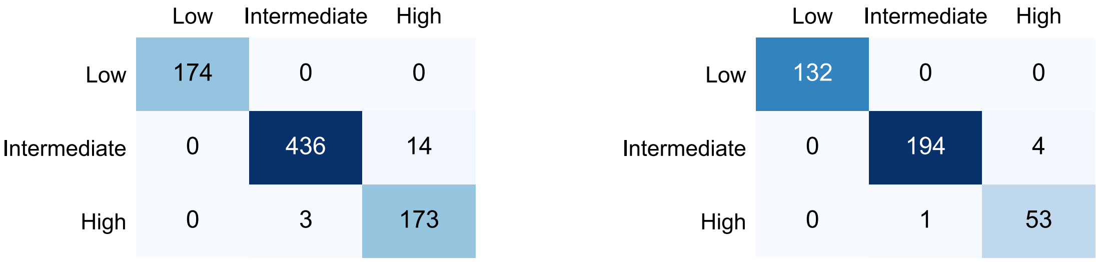
Fig. 3. Both the datasets were split with a ratio 80:20 to create train and validation sets. 'train loss' shows the loss of the trained model on the training data. 'val loss' shows the loss of the trained model on the validation data. Training loss and validation loss decrease and become stable after a certain point which reveals an optimally fit trained model. 'Full data loss' shows the loss plot of the 100% data.



(a) Loss plots of multiple training either using individual **A**1, **A**2, **A**3 matrix or using combined alpha fused matrix.

(b) Loss plots of multiple training done with varied α values. $\alpha = 0.7$ gave the lowest loss and early saturation.

Fig. 4. Loss plots obtained with different settings of **A1**, **A2**, **A3** matrix and varying α.



(a) On MMRF

(b) On MMIn

Fig. 5. Confusion Matrices obtained from the GCRS model trained for risk-stage prediction.

3.1. Ablation studies

In this section, we performed an ablation study using individual **A1**, **A2**, and **A3** matrices to assess the performance of our proposed alpha weighted fusion matrix. To this end, we trained the proposed GCRS model with respect to these three individual matrices separately. **A1**, **A2**, and **A3** matrices are obtained using ranks from multivariate cox hazard analysis, univariate cox hazard ratios, and p-values from log-rank on Kaplan Meier analysis, respectively. Fig. 4(a) shows the training loss learning curves with respect to the number of epochs. As seen from this figure, our proposed alpha-weighted fusion approach has minimal loss compared to the other trained models with individual matrices. This indicates optimal training with the proposed method in terms of early loss saturation and an increase in the performance of risk stratification compared to the existing methods.

Our alpha fusion methodology depends on a single hyperparameter, α. Our grid search experiment helped us find the most optimal α value. To see how different alpha values affect GCRS model performance, we plot the training loss values with respect to the number of epochs for different alpha values in the range of 0.1 to 0.9 in Fig. 4(b). It is clear from this figure that a value of $\alpha = 0.7$ leads to optimal model training in terms of lesser loss value compared to that obtained with the other α values. Moreover, as can be seen from Fig. 4(a), adjacency matrix **A3** results in better loss saturation compared to the other two matrices, leading to lesser weightage to be given to this matrix while training the model. In other words, a higher weight for multivariate cox hazard analysis merely suggests that the information in matrices **A1** and **A2** are required to be emphasized more to increase their contribution to the model's learning.

3.2. Statistical analysis on the parameters used in GCRS

The underlying topological differences and the differences in clinical and laboratory features of patients among the three risk groups (low, intermediate, and high) are investigated statistically by applying the unpaired Wilcoxon rank-sum test. Kruskal–Wallis test is also used to compare the three risk groups together. Furthermore, the baseline clinical and laboratory features of patients from all three risk groups are visualized using boxplots. To do this, we first stratified patients into the three risk groups using the trained GCRS model as GCRS-1, GCRS-2, and GCRS-3 for low-risk, intermediate risk, and high-risk groups, respectively. Boxplot of each parameter is plotted in Fig. 6 for the MMRF dataset and in Fig. 7 for the MMIn dataset. We observed an increasing trend of age and β2m parameters with increase in risk. In contrast, as the risk increases, a decreasing trend is observed in albumin, eGFR, and hemoglobin parameters.

On comparing intermediate and low-risk groups (Fig. 6), we observe that the intermediate risk group patients belong to higher age (p = 1.30e−05), have decreased parameters such as albumin (p = 1.09e−21), calcium (p = 1.29e−02), eGFR (p = 2.64e−04), hemoglobin (p = 2.32e−20) and have increased β2m parameter (p = 9.47e−16). Similarly, looking at the comparison between high and intermediate risk groups, the high-risk groups patients belong to higher age (p = 2.44e−06), have decreased parameters such as albumin (p = 1.09e−07), calcium (p = 7.70e−02), eGFR (p = 2.97e−40), hemoglobin (p = 4.40e−28) and have increased β2m parameter (p = 1.06e−62). This indicates that the median values of all the parameters differ significantly among different pairs of the three risk groups (p<0.05), thereby, substantiating that the three risk populations identified are indeed different. All parameters were also found to be statistically different (at p<0.05) across the three risk groups using the Kruskal–Wallis test.

Given the potential differences among all the three groups in the MMRF dataset, further investigation was made on the MMIn dataset, as shown in Fig. 7. Again, for all the three risk groups, low-, intermediate- and high-risk groups, the paired Wilcoxon rank-sum test for all the possible pairs of parameters were found to be statistically significantly different with p<0.05. A similar result is obtained using the Kruskal–Wallis test among the three risk group pairs.

3.3. Comparison with previous methods

The risk-group labels identified by the proposed GCRS methods were analyzed for OS and PFS and compared with those obtained with the ISS, R-ISS, MRS, and CRSS methods on the MMRF and MMIn datasets. To the best of our knowledge, no deep learning-based risk staging method has been proposed in MM so far to compare with. Moreover, graph-based methods are one of the most recent and popular methods in the deep learning domain. Hence, this approach has been used in this work. This is to note that ISS and MRS do not utilize the genetic feature, while R-ISS, CRSS, and the proposed GCRS also use the cytogenetic feature. So far, the staging methods of R-ISS and CRSS using the cytogenetic feature have been observed to perform better than ISS and MRS that do not utilize the cytogenetic feature. GCRS, besides using the cytogenetic feature, also leverages the new generation of graph-based deep learning methodology and hence, is hypothesized to generate better results compared to the existing methods in terms of better statistical separability of the three risk groups.

Results on the MMRF dataset: KM survival analysis of GCRS groups in MMRF dataset (Fig. 8) provided a statistically significant difference in OS between GCRS-1 and GCRS-2 groups with p = 0.00024. This performance is better than the ISS method and compared to the other three methods, with p = 0.0001 between CRSS-1 and CRSS-2, with p = 6.92e–15 between MRS-1 and MRS-2, with p = 0.03 between R-ISS1 and R-ISS2, and with p = 0.001 between ISS-1 and ISS-2. This performance improved considerably between GCRS-2 and GCRS-3 groups with p = 6e–14 compared to the other methods with p = 3.5e–09 between CRSS-2 and CRSS-3, p = 1.13e–05 between MRS-2 and MRS-3, p = 5.2e–06 between R-ISS2 and R-ISS3, and p = 0.0002 between ISS-2 and ISS-3.

Table 3
p-values for the Kaplan–Meier curves from MMRF predicted labels from different models indicating separability between the three risk stages.

Model	Overall survival	Progression free survival
ISS	1.56e–10	2.49e–09
R-ISS	6.58e–08	1.74e–05
MRS	1.86e–14	1.15e–06
CRSS	1.09e–15	8.65e–12
GCRS	**2.35e–17**	**2.32e–13**

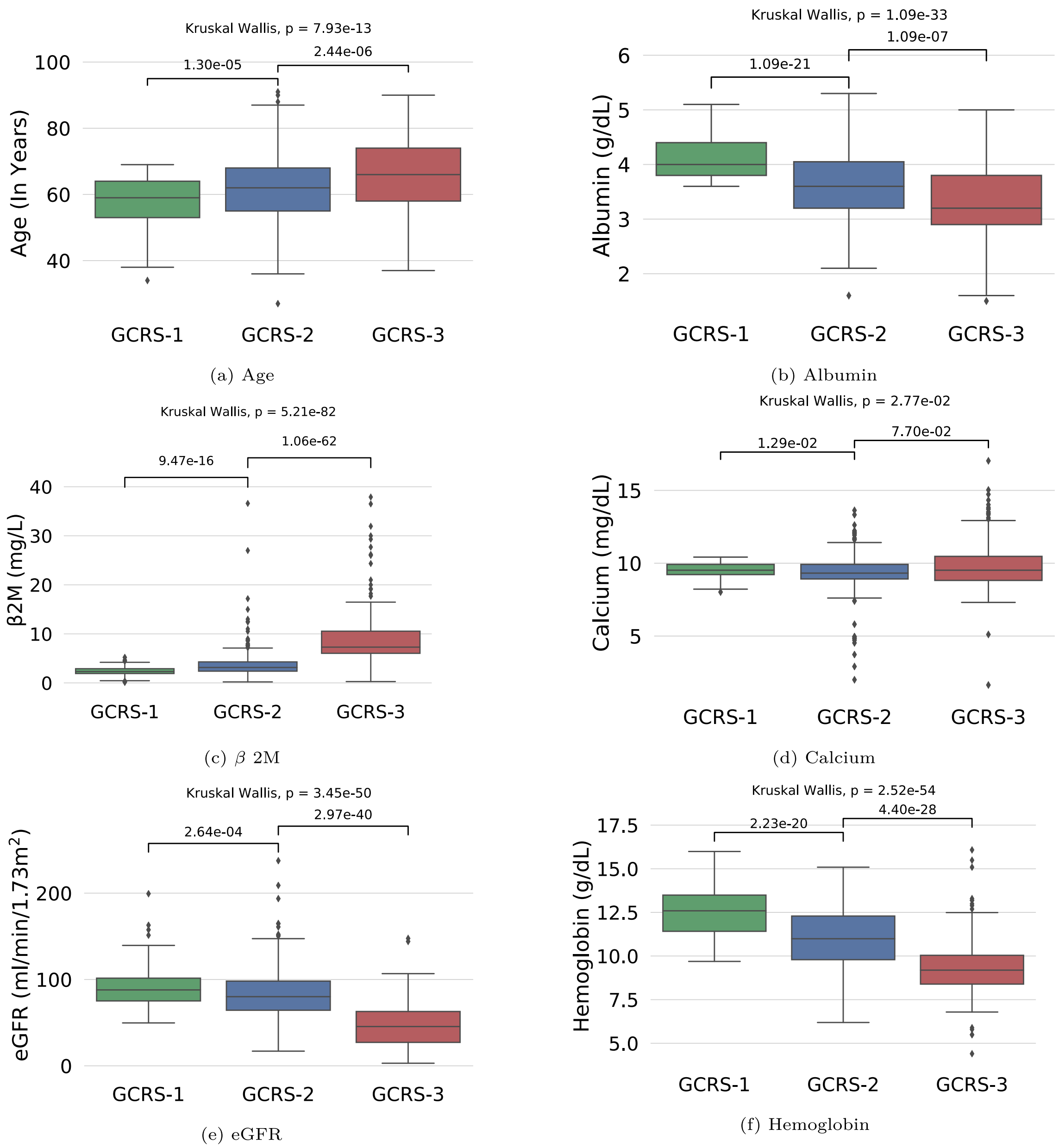
KM survival analysis of GCRS groups (Fig. 9) further revealed statistically significant difference in PFS between GCRS-1 and GCRS-2 with p = 0.001, between CRSS-1 and CRSS-2 with p = 0.0006, between MRS-1 and MRS-2 with p = 6.4e–12, between R-ISS1 and R-ISS2 with p = 0.01, and between ISS-1 and ISS-2 with p = 0.001. Improved Statistical significant difference was found between GCRS-2 and GCRS-3 groups with p = 4.4e–12 compared to the other methods, with p = 4.4e–07 between CRSS-2 and CRSS-3, p = 6.3e–08 between MRS-2 and MRS-3, p = 0.0006 between R-ISS2 and R-ISS3 and p = 0.001 between ISS-2 and ISS-3. Furthermore, overall p-values of KM survival analysis in MMRF dataset showed more separability in GCRS risk grouping (p < 2.35e–17 for OS and p < 2.32e–13 for PFS) compared to others as shown in Table 3.

Results of multivariate Cox hazard analysis also indicated a superior performance of our model, wherein a higher C-Index was achieved with GCRS (0.684 for OS and 0.62 for PFS) (See Table 4). With GCRS, we obtained the C-Statistic of 0.684 (HR = 2.99) for OS and 0.62 (HR = 1.84) for PFS on the MMRF data as compared to 0.676 (HR = 2.85) and 0.61 (HR = 1.79) with CRSS; 0.62 (HR = 2.11) and 0.60 (HR = 1.60) with MRS; 0.62 (HR = 2.26) and 0.58 (HR = 1.61) with R-ISS; and 0.662 (HR = 1.95) and 0.60 (HR = 1.51) with ISS. The risk of progression and mortality increased for GCRS 2vs1 and GCRS 3vs1.

Results on MMIn dataset: Similarly, KM survival analysis of GCRS groups (Fig. 10) on MMIn dataset indicated statistically significant difference in OS between GCRS-1 & GCRS-2 groups with p = 1.04e–07, between CRSS-1 & CRSS-2 with p = 1.1e–08, between MRS-1 & MRS-2 with p = 5.91e–09, between R-ISS1 & R-ISS2 with p = 0.04, and between ISS-1 & ISS-2 with p = 0.28. Improved statistically significant difference was found between GCRS-2 & GCRS-3 groups with p = 2.1e–14 compared to p = 0.02 between CRSS-2 & CRSS-3, p = 0.001 between MRS-2 & MRS-3, p = 1.9e–05 between R-ISS2 & R-ISS3, and p = 0.0005 between ISS-2 & ISS-3.

KM survival analysis of GCRS groups (Fig. 11) in MMIn further revealed a statistically significant difference in PFS between GCRS-1 and GCRS-2 with p = 0.003. This performance was comparable to that of CRSS and MRS with p = 0.0003 between CRSS-1 and CRSS-2 and p = 0.00125 between MRS-1 and MRS-2, and better than R-ISS and ISS with p = 0.3 between R-ISS1 and R-ISS2, and p = 0.7 between ISS-1 and ISS-2. Improved statistically significant difference was found between GCRS-2 and GCRS-3 groups with p = 5.4e–09, compared to p = 0.003 between CRSS-2 and CRSS-3, p = 0.0054 between MRS-2 and MRS-3, p = 0.01 between R-ISS2 and R-ISS3, and p = 0.001 between ISS-2 and ISS-3. Overall, p-values of KM survival analysis in both the datasets showed better separability in GCRS (for OS p < 3.02e–12, and for PFS p < 6.00e–07) risk grouping compared to the other methods as shown in Table 5.

Multivariate Cox Hazard analysis also indicated superior or comparable performance of GCRS in terms of C-Index on the MMIn dataset (See Table 6). The C-Statistic with GCRS on the MMIn data was 0.676



(a) Age

(b) Albumin

(c) β 2M

(d) Calcium

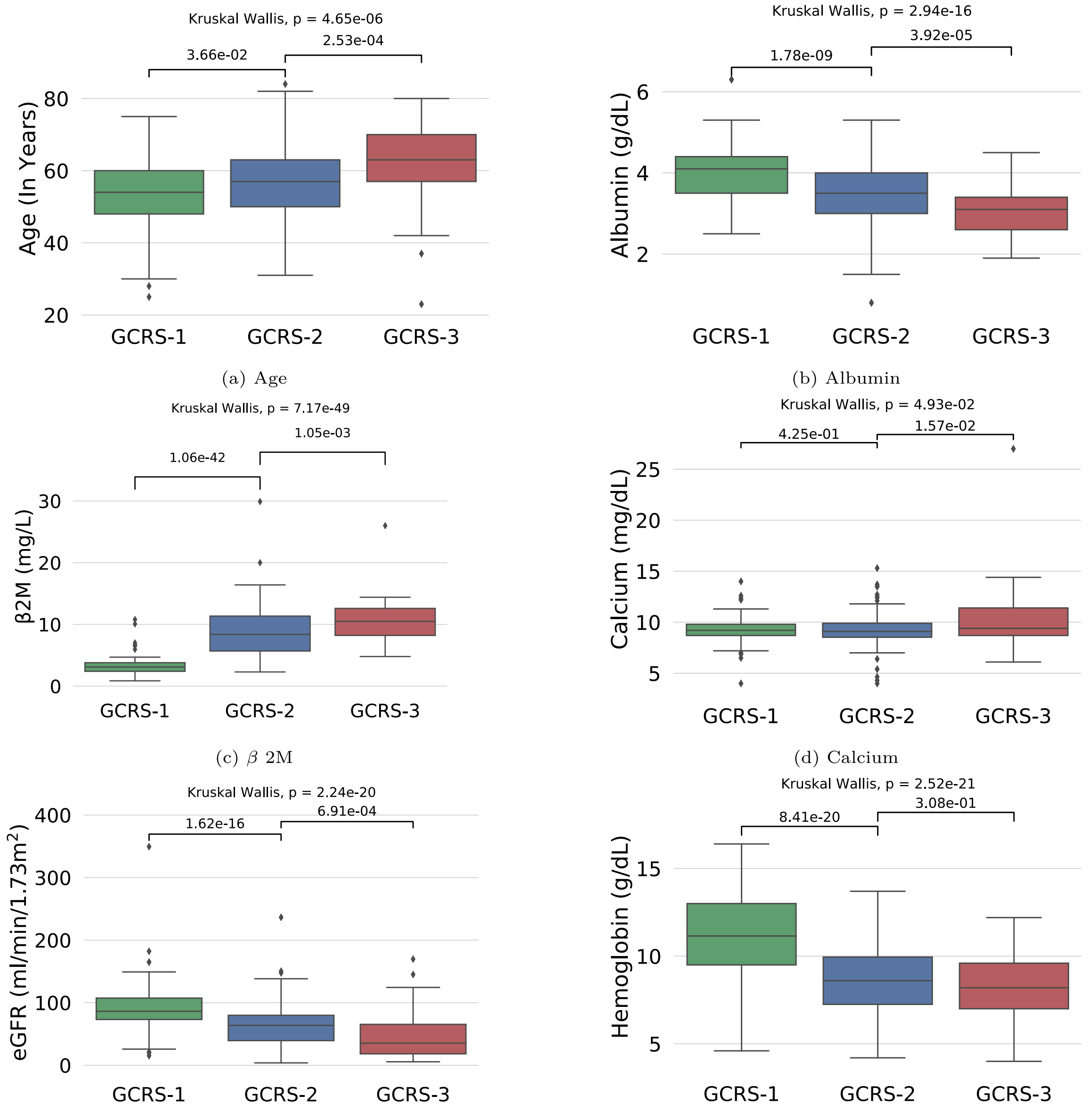
(e) eGFR

(f) Hemoglobin

Fig. 6. Boxplots showing the variation of six parameters after assigning patients of MMRF dataset to GCRS-1, GCRS-2, GCRS-3 risk groups.

Table 4
Cox Hazard Ratios on MMRF dataset for risk stages predicted from different models.

Model		Overall survival			Progression free survival		
		Hazard Ratio	p-Value	C-Index	Hazard Ratio	p-Value	C-Index
	2vs1	1.99	2.25e−3		1.52	1.52e−3	
ISS	3vs1	3.84	<5e−06	0.662	2.28	<5e−06	0.60
	Overall	1.95	<5e−06		1.51	<5e−06	
	2vs1	1.79	0.03		1.49	0.015	
R-ISS	3vs1	4.66	<5e−06	0.618	2.6	1e−5	0.578
	Overall	2.26	<5e−06		1.61	0.00001	
	2vs1	2.40	0.00002		1.88	<5e−06	
MRS	3vs1	4.62	<5e−06	0.65	2.55	<5e−06	0.60
	Overall	2.11	<5e−06		1.6	<5e−06	
	2vs1	4.10	3.4e−4		1.76	8.10e−04	
CRSS	3vs1	10.61	<5e−06	0.676	3.19	<5e−06	0.61
	Overall	2.85	<5e−06		1.79	<5e−06	
	2vs1	3.88	6.1e−4		1.70	1.6e−3	
GCRS	3vs1	10.8	<5e−06	**0.684**	3.28	<5e−06	**0.62**
	Overall	2.99	<5e−06		1.84	<5e−06	



(a) Age

(b) Albumin

(c) β 2M

(d) Calcium

(e) eGFR

(f) Hemoglobin

Fig. 7. Boxplots showing the variation of six parameters after assigning patients of MMIn dataset to GCRS-1, GCRS-2, GCRS-3 risk groups.

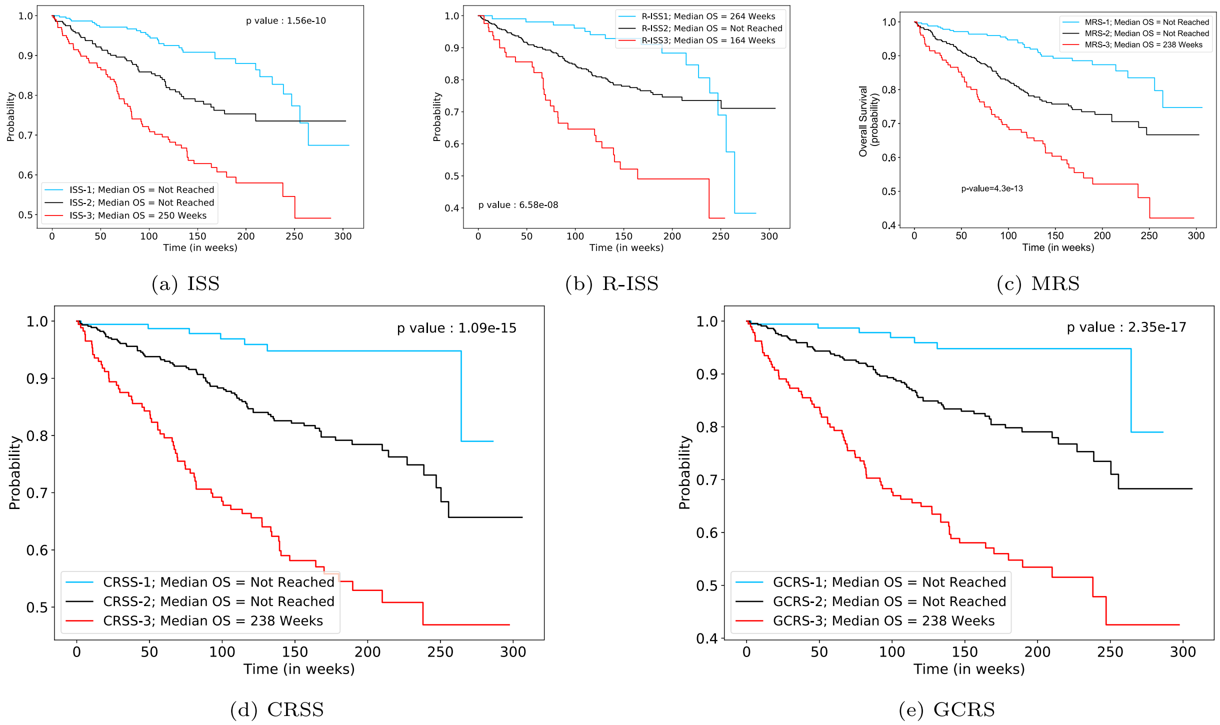
(HR = 2.69) for OS and 0.6 (HR = 1.71) for PFS compared to 0.676 (HR = 2.43) and 0.61 (HR = 1.8) with CRSS, 0.63 (HR = 1.89) and 0.57 (HR = 1.37) with MRS, 0.63 (HR = 2.32) and 0.57 (HR = 1.42) with R-ISS, and 0.60 (HR = 1.87) and 0.57 (HR = 1.41) with ISS. The progression and mortality risk increased for GCRS 2vs1 and GCRS 3vs1.

3.4. Model interpretation

SHapley Additive exPlanations (SHAP) analysis was carried out to observe the impact of individual parameters on risk stages predicted by GCRS. SHAP is a game-theoretic approach that is used to interpret the decisions of a machine learning model. We employed the use of shapely values calculated from the gradients for each parameter in our deep neural network (See Figs. 12 and 13). This was achieved using SHAP's Gradient Explainer [21]. Key contributors to risk stage predictions in the MMRF dataset were age, β2m, and eGFR. It can also be visualized from Figs. 12(a), and 12(c) that higher β2m values are indicative of high tumor growth, and its higher value positively impacts the decision to the high-risk stage of GCRS-3. A lower value of β2m positively impacts the decision to the low-risk stage of GCRS-1. Similarly, we observe that a higher value of eGFR, which is an indicative of good kidney function, positively impacts the decision to GCRS-1. This implies that a patient with high eGFR value belongs to the low-risk stage. Further, the key contributors of risk stage predictions in the MMIn dataset were ordered as eGFR, β2m, and age. A similar observation could be made from Figs. 13(a) and 13(c) that a higher eGFR, lower β2m, and a lower value of age positively contribute to the low-risk stage (GCRS-1) and negatively contribute towards the high-risk stage (GCRS-3).

4. Discussion

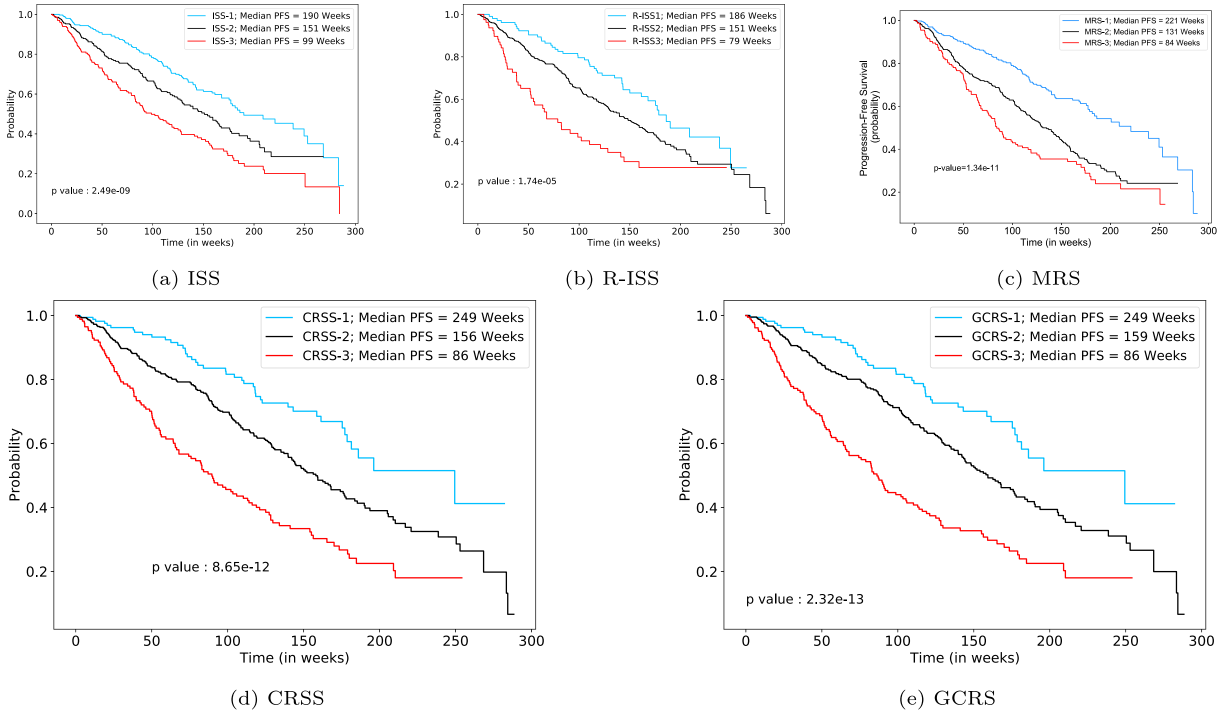
The contributions of deep learning to various domains have been of paramount importance. This is especially true in medical that often encompasses big and/or non-euclidean data. Graph convolutional neural networks, an innovation in deep learning, run natively on graph topology structures that can capture non-euclidean data. It can be intuitively understood that patients with similar diagnoses would be



(a) ISS (b) R-ISS (c) MRS

(d) CRSS (e) GCRS

Fig. 8. Overall Survival: Kaplan–Meier Curves on MMRF predicted labels from (a) ISS (b) R-ISS (c)MRS (d) CRSS and the proposed (e) GCRS. *p*-Values indicate the separability between risk stages, where 1 denotes low-risk, 2 denotes Intermediate Risk, and 3 denotes the high-risk category.

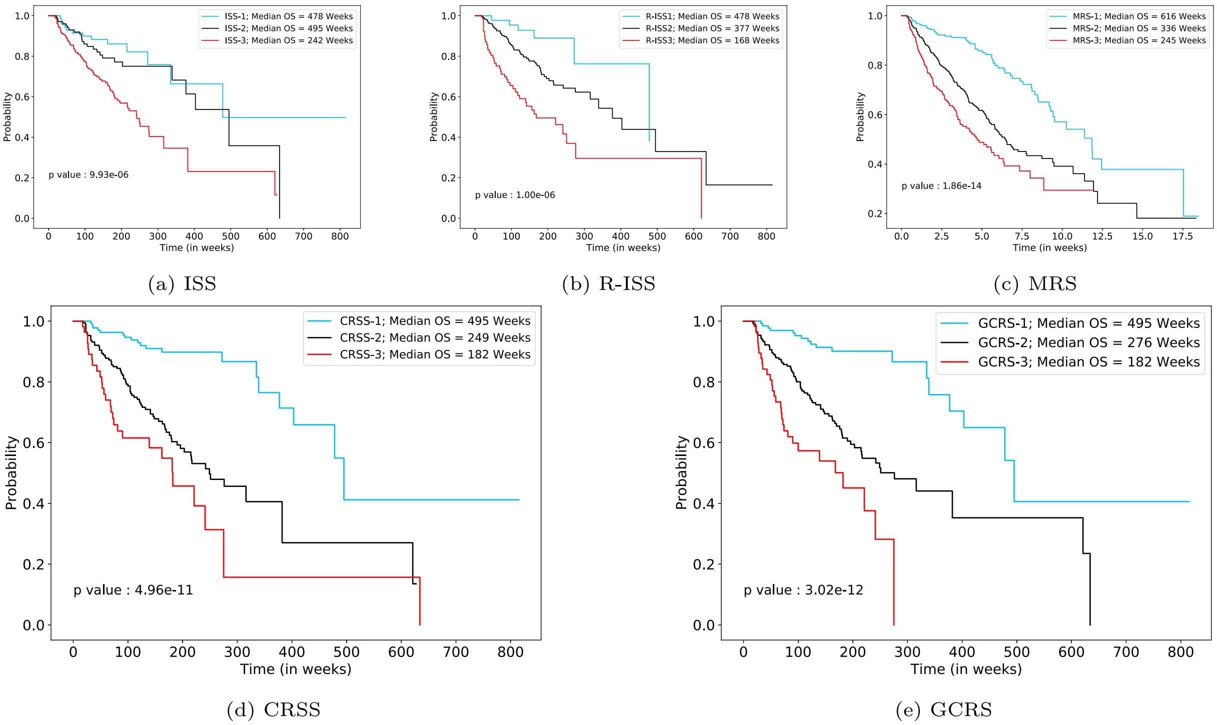


(a) ISS (b) R-ISS (c) MRS

(d) CRSS (e) GCRS

Fig. 9. Progression Free Survival: Kaplan–Meier Curves on MMRF predicted labels from (a) ISS (b) R-ISS (c) MRS (d) CRSS and proposed (e) GCRS. *p*-Values indicate separability between the risk stages, where 1 denotes low-risk, 2 denotes Intermediate Risk, and 3 denotes the high-risk category.

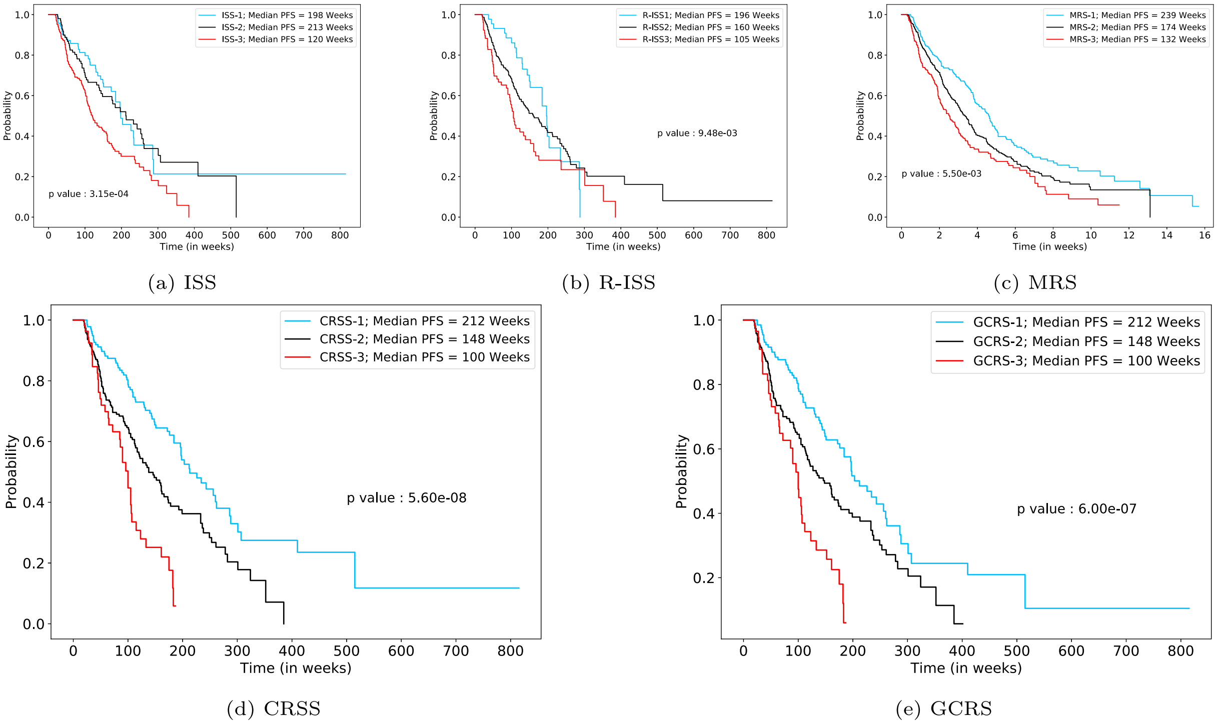
connected in this graph structure. Hence, it is a natural structure that can be exploited for comparatively rich information learning and extraction for better risk staging in newly diagnosed multiple myeloma cancer patients. Multiple myeloma is a cancer of plasma cells where the overall survival varies from several months to more than ten years [33]. The neoplastic proliferation of plasma cells in the bone marrow and consequent release of osteolytic cytokines and monoclonal protein are responsible for clinical manifestations such as bone pains and hypercalcemia due to increased bone turnover, compromised renal function due to renal damage caused by monoclonal protein and anaemia due to reduced haemoglobin. Therefore, the degree of derangement of serum calcium, haemoglobin and creatinine, which are reflective of increased bone turnover, anaemia and renal function, respectively, influences the performance status of the patient and also impacts the clinical outcome.



(a) ISS (b) R-ISS (c) MRS

(d) CRSS (e) GCRS

Fig. 10. Overall Survival: Kaplan–Meier Curves on MMIn predicted labels from (a) ISS (b) R-ISS (c) MRS (d)CRSS and proposed (e) GCRS. *p*-Values indicate the separability between risk stages, where 1 denotes low-risk, 2 denotes Intermediate Risk, and 3 denotes high-risk category.

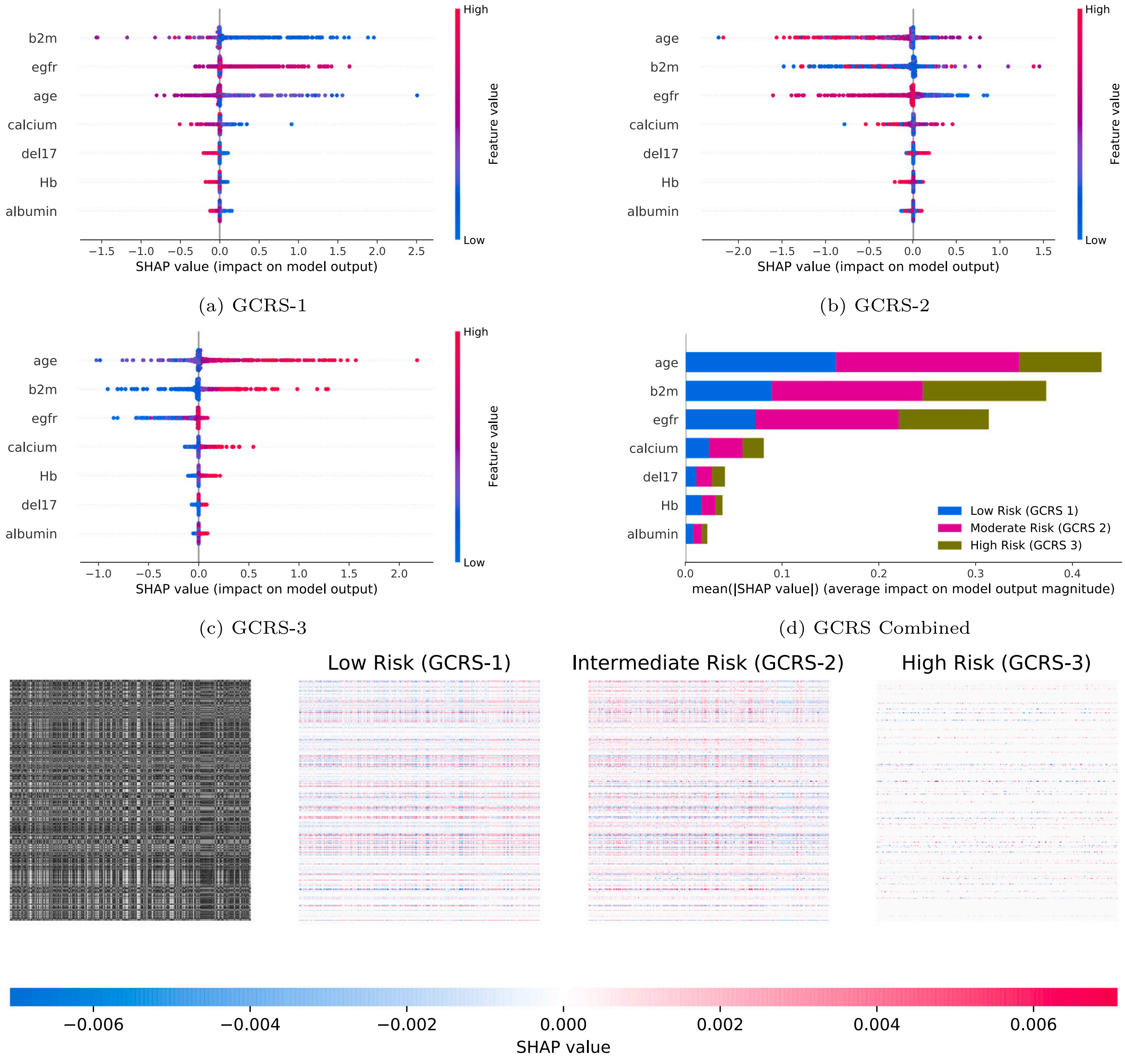


(a) ISS (b) R-ISS (c) MRS

(d) CRSS (e) GCRS

Fig. 11. Progression Free Survival: Kaplan–Meier Curves on MMIn predicted labels from (a) ISS (b) R-ISS (c) MRS (d) CRSS and proposed (e) GCRS. *p*-Values indicate the separability between risk stages, where 1 denotes low-risk, 2 denotes Intermediate Risk, and 3 denotes high-risk category.

Although tremendous efforts have been made over the past few decades to improve the diagnosis and treatment of MM, it remains incurable because of its heterogeneity. Therefore, an efficient MM-stratification system is needed to separate high-risk patients at diagnosis, providing optimal therapy to these high-risk patients and ultimately prolonging their overall survival. R-ISS is the current standard of care for predicting risk stages in myeloma patients. This risk staging is based on only a few prognostic parameters such as β2m, albumin, LDH, and cytogenetic abnormalities. Several published studies have demonstrated the robustness of serum levels of albumin, β2m, renal functions (indicated by eGFR), and high-risk genetic aberrations (HRCA) for risk stratification in MM. Considering the heterogeneity in MM patients



(e) Adjacency

Fig. 12. Shap values from our GCRS Model trained on the MMRF dataset. (a), (b) and (c) depict the impact of each parameter on a specific risk stage; (d) visualizes the overall mean impact of each parameter on each risk stage; and (e) highlights (in red) all those patients in the adjacency matrix who have a higher impact on the individual risk stage learning of the model.

Table 5
p-values for the Kaplan–Meier curves from MMIn predicted labels from different models indicating the separability between the three risk stages.

Model	Overall survival	Progression free survival
ISS	9.93e−06	3.15e−4
R-ISS	1.00e−06	9.48e−3
MRS	**1.86e−14**	1.14e−06
CRSS	4.96e−11	**5.60e−08**
GCRS	**3.02e−12**	6.00e−07

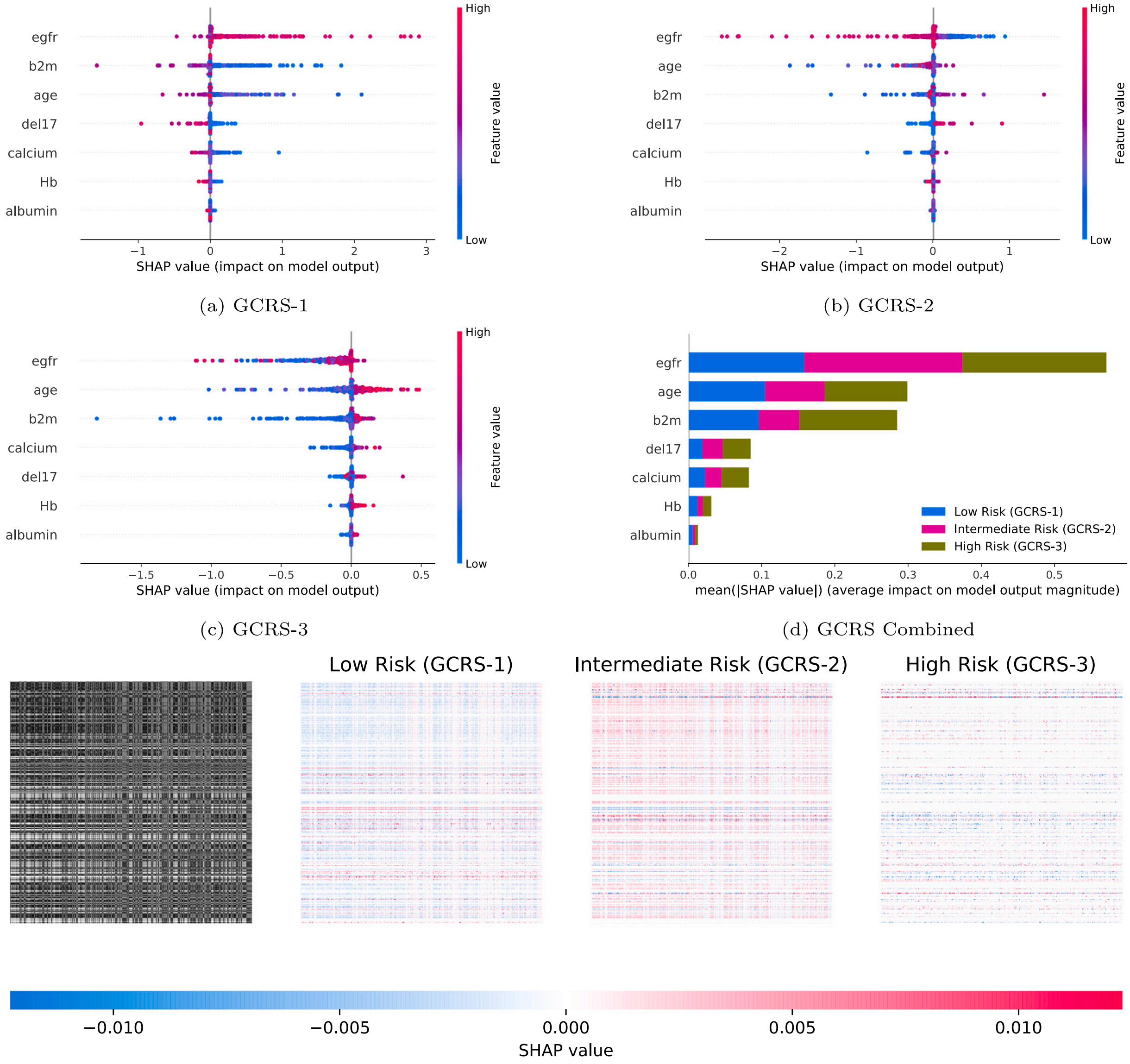
globally, it is desirable to have a risk-staging system based on multiple known adverse clinical and molecular prognostic factors. Recently, AI-supported risk-staging models, MRS [14] and CRSS [15] have been developed that utilize age, albumin, β2m, calcium, hemoglobin, and eGFR to predict risk stages in MM patients. Age, hemoglobin, and calcium levels were found to influence the categorization of patients into different risk groups in MRS [14] and CRSS [15] and hence, are evaluated in this study. Using these parameters for a more comprehensive implementation of the risk stratification system in clinical practice is imperative. Further, given the efficacy of machine learning methods in risk stage prediction, we have presented a graph convolutional neural network-based risk-staging method for MM in this work.

4.1. Parameters characteristics of different risk groups

The credibility of the proposed GCRS model was validated by Kruskal–Wallis test and Wilcoxon rank-sum test. Kruskal–Wallis test was used to evaluate the significance of the difference in the median values among the three groups. Wilcoxon rank-sum test is similar to the Kruskal–Wallis test, but it is computed between two groups. Figs. 6 and 7 revealed statistically significant variations in the median values of all the parameters across the three groups with $p < 0.05$. Similar results are observed with the Wilcoxon rank-sum test, which revealed statistically significant variations ($p < 0.05$) in the median values of the parameters between two successive risk groups (GCRS-1 and GCRS-2; GCRS-2 and GCRS-3). Further, the risk of developing MM increases

Table 6
Cox Hazard Ratios for MMIn predicted risk stages from different models.

Model		Overall survival			Progression free survival		
		Hazard Ratio	p-Value	C-Index	Hazard Ratio	p-Value	C-Index
ISS	2vs1	1.41	0.3	0.60	1.08	0.7	0.57
	3vs1	3.12	1.7e−4		1.84	1.56e−3	
	Overall	1.87	0.00001		1.41	0.00024	
R-ISS	2vs1	2.31	0.04	0.636	1.24	0.33	0.57
	3vs1	5.37	1.3e−4		1.92	0.009	
	Overall	2.32	<5e−06		1.42	0.00418	
MRS	2vs1	2.29	<5e−06	0.63	1.36	0.00157	0.57
	3vs1	1.23	<5e−06		1.79	<5e−06	
	Overall	1.89	<5e−06		1.37	<5e−06	
CRSS	2vs1	3.95	<5e−06	0.67	1.76	3.00e−04	0.6
	3vs1	6.43	<5e−06		3.27	<5e−06	
	Overall	2.43	<5e−06		1.8	<5e−06	
GCRS	2vs1	3.74	<5e−06	**0.676**	1.59	2.95e−3	**0.6**
	3vs1	7.79	<5e−06		3.0	<5e−06	
	Overall	2.69	<5e−06		1.71	<5e−06	



(a) GCRS-1

(b) GCRS-2

(c) GCRS-3

(d) GCRS Combined

(e) Adjacency

Fig. 13. Shap values from our GCRS Model on the MMIn dataset. (a), (b) and (c) depict the impact of each parameter on a specific risk stage; (d) visualizes the overall mean impact of each parameter on each risk stage; and (e) highlights (in red) all those patients in the adjacency matrix who have a higher impact on the individual risk stage learning of the model.

Table 7
Advantages and disadvantages of different methods used for risk stage prediction in MM.

	Advantages	Disadvantages
ISS	—Relies only on two parameters, albumin and β2m for predicting risk stage in MM.	—Does not utilize high risk cytogenetic abnormalities information for risk stage prediction. —Ethnicity information is not used for risk stage prediction in ISS. —No machine learning method has been employed in ISS.
R-ISS	—Includes high risk cytogenetic abnormalities for risk stage prediction along with albumin, β2m, LDH.	—Ethnicity information is also not used for risk stage prediction in R-ISS. —No machine learning method has been employed in R-ISS.
MRS	—Makes use of easy-to-acquire parameters for risk stage prediction like age, albumin, β2m, calcium, hemoglobin and eGFR. —Machine learning methods like BIRCH clustering and j48 decision tree classifier have used for building MRS. —Suitable for settings where genomic tests cannot be performed owing to geographical or economic constraints. —Performs better than ISS and comparable to R-ISS even without utilizing high risk cytogenetic abnormalities information.	—Ethnicity information is not used for risk stage prediction in MRS. —Validated on dataset which is roughly 25% of the size of the dataset used for building R-ISS.
CRSS	—Utilizes ethnicity information and high risk cytogenetic information along with parameters used in MRS. —Machine learning methods like GMM clustering, Agglomerative clustering and decision tree classifier has been used for building CRSS. —Performs better than ISS, R-ISS and MRS with higher values of C-index and hazard ratios and lower p-values.	—Validated on dataset which is roughly 25% of the size of the dataset used for building R-ISS.
GCRS	—Utilizes ethnicity information and high risk cytogenetic information along with parameters used in MRS. —Graph convlutional network based deep learning method has been used for building GCRS. —Performs better than ISS, R-ISS, MRS and CRSS with higher values of C-index and hazard ratios and lower p-values.	—Validated on dataset which is roughly 25% of the size of the dataset used for building R-ISS.

with older age, elevated levels of serum β2m, and calcium and lower levels of albumin, hemoglobin, and eGFR. A similar observation was noticed in the boxplots, where the high-risk stage was associated with older age, raised levels of serum β2m and calcium, and lower levels of the rest of the parameters. On the contrary, in the low-risk stage, the prognostic parameters slightly deviated from their normal levels. This is justified because the low-risk stage implies that cancer is still in the initial stage without causing a profound negative impact. Thus, these findings pertain to the observations in MM patients.

Furthermore, GCRS models built for the MMIn and MMRF datasets were interpreted using SHAP to demonstrate the relevance of the predicted risk stages. For the MMRF data, elevated levels of β2m and older age contributed to high risk in myeloma patients. In contrast, lower levels of β2m and high levels of eGFR contributed to the low-risk stage, as shown in Fig. 12. For the MMIn data, elevated levels of β2m and older age contributed to high-risk, whereas high levels of eGFR and lower levels of β2m contributed to the low-risk stage as shown in Fig. 13. These observations are in agreement with the information known for MM patients. In addition, it was observed that the order of impact of age was the highest for risk stage prediction in the MMRF dataset compared with the MMIn dataset, where the order of impact of eGFR was the highest. The difference in the rankings can be attributed to the varying ethnicities.

4.2. Performance of GCRS as compared to CRSS, MRS, R-ISS and ISS

On both the MMRF and MMIn datasets, the proposed GCRS performed better overall than CRSS, MRS, ISS, and R-ISS for the prediction of OS and PFS. We compared performances in terms of C-index, HR, and *p*-values. The performance w.r.t. HR was comparable between GCRS and CRSS. In the MMRF dataset, 800 patients out of 900 were given GCRS labels. Of these patients, 174 (21.75%) were labeled as GCRS-1, 439 (54.86%) were labeled as GCRS-2, and 187 (23.38%) were labeled as GCRS-3. In the MMIn dataset, 384 patients out of 1070 were given GCRS labels. Of these patients, 132 (34.38%) were labeled as GCRS-1, 195 (50.78%) were labeled as GCRS-2, and 57 (14.84%) were labeled as GCRS-3. Table 7 provides a comparison of the advantages and disadvantages of the different methods of risk stage prediction in MM.

4.3. Scope, limitations and future work

The present study highlights the advantages of the availability of large annotated datasets with treatment outcomes that allows for the development of advanced statistical tools such as data augmentation, data imputation, and self-trainable models, required for building robust risk prediction models. The most robust and weighted parameters can be inferred and mathematically/sequentially introduced in cancer risk stratification. Nevertheless, this work has some limitations as well. The proposed deep learning-enabled model utilizes HRCA information in addition to multiple prognostic factors. Since HRCA information was unavailable for all patients, our data size was reduced significantly. However, we plan to address this limitation by deploying an online trainable GCRS application. It will collect data from independent groups if the application users agree, which could be further used for training and validation. Moreover, we would also work on optimizing the patient network for more sparsity. This could be achieved by making the adjacency graph a trainable parameter in the network optimization problem. The current study addressed an upfront risk prediction, which is a static assessment based on baseline clinical parameters. An interesting direction can be to predict and track the risk stage during the course of treatment, also known as the ‘Dynamic Risk Stratification’.

The response to treatment is influenced by patient-related factors, type and sequencing of therapies, and disease-related factors. Although disease-related factors have been evaluated in the present study, an adequate assessment of all the patient-related factors is not possible in the biological domain and hence, in this disease. The response to therapy and reversal of clinical manifestations with anti-myeloma treatment are some indirect measurand of the same. It has been shown that the first response after induction therapy and, the depth of response after chemotherapy as well as after autologous stem cell transplantation influence the long term treatment outcomes in MM [34–36]. A recent study has reported that patients demonstrating reversal of renal dysfunction after anti-myeloma therapy have better survival than those who have sustained renal dysfunction [37]. Recent studies have demonstrated evolution of myeloma genome after chemotherapy, which can modify risk predictions [1,38]. In future, risk modeling can be expanded to include the impact of different combination and number of therapies given to a patient paving way for personalized medicine in MM. Overall, it would be interesting to compare the upfront static risk stratification with the dynamic risk modeling to evaluate the impact and contribution of patient related factors and treatments on clinical outcomes.

5. Conclusion

In this work, we proposed a superior deep learning-assisted risk prediction system, GCRS, for MM patients belonging to the Indian and American populations. GCRS leverages a multi-spatial patient network that assumes that patients with a similar survival analysis statistic would be closer and connected to a graph topology. Separate risk-staging models were built for both the MMRF and MMIn datasets. Models were trained using easily acquirable laboratory and clinical parameters: age, albumin, β2m, calcium, eGFR, hemoglobin, and the HRCA information. Risk stratification achieved by deep learning enabled GCRS can better separate the patients into different risk groups as compared with the CRSS. Higher concordance indices and hazard ratios reveal superior performance of GCRS. Furthermore, the clinical and biological significance of each parameter in determining the risk stage in MM was deduced via SHAP analysis on both datasets. Our study also highlights the importance of deploying deep learning in building GCRS, thereby enhancing the prediction of survival outcomes and separability of risk stages in MM patients. For future work, we suggest formulating an online trainable GCRS application that could be trained and validated on the data collected from independent sources. MM is a highly heterogeneous cancer. Hence, inclusion of a dataset belonging to multiple ethnicity groups can facilitate the generation of an efficient model for risk stratification with excellent global utility.

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.

Acknowledgments

The authors acknowledge Multiple Myeloma Research Foundation (MMRF) , United States for providing the dataset. These data were generated as part of MMRF Personalized Medicine Initiative. The authors would also like to acknowledge the Infosys Center for Artificial Intelligence, IIIT-Delhi, India, for the financial support (Grant No. SRP-205) in this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106048.

References

[1] A. Farswan, L. Jena, G. Kaur, A. Gupta, R. Gupta, L. Rani, A. Sharma, L. Kumar, Branching clonal evolution patterns predominate mutational landscape in multiple myeloma, Am. J. Cancer Res. 11 (11) (2021) 5659.

[2] B.G. Durie, S.E. Salmon, A clinical staging system for multiple myeloma correlation of measured myeloma cell mass with presenting clinical features, response to treatment, and survival, Cancer 36 (3) (1975) 842–854.

[3] P.R. Greipp, J.S. Miguel, B.G. Durie, J.J. Crowley, B. Barlogie, J. Bladé, M. Boccadoro, J.A. Child, H. Avet-Loiseau, R.A. Kyle, et al., International staging system for multiple myeloma, J. Clin. Oncol. 23 (15) (2005) 3412–3420.

[4] A. Palumbo, H. Avet-Loiseau, S. Oliva, H.M. Lokhorst, H. Goldschmidt, L. Rosinol, P. Richardson, S. Caltagirone, J.J. Lahuerta, T. Facon, et al., Revised international staging system for multiple myeloma: a report from international myeloma working group, J. Clin. Oncol. 33 (26) (2015) 2863.

[5] A. Rago, S. Grammatico, T. Za, A. Levi, S. Mecarocci, A. Siniscalchi, L. De Rosa, S. Felici, V. Bongarzoni, A.L. Piccioni, et al., Prognostic factors associated with progression of smoldering multiple myeloma to symptomatic form, Cancer 118 (22) (2012) 5544–5549.

[6] M. Schinke, G. Ihorst, J. Duyster, R. Wäsch, M. Schumacher, M. Engelhardt, Risk of disease recurrence and survival in patients with multiple myeloma: A german study group analysis using a conditional survival approach with long-term follow-up of 815 patients, Cancer 126 (15) (2020) 3504–3515.

[7] M. Dimopoulos, S. Delimpasi, E. Katodritou, A. Vassou, M. Kyrtsonis, P. Repousis, Z. Kartasis, A. Parcharidou, M. Michael, E. Michalis, et al., Significant improvement in the survival of patients with multiple myeloma presenting with severe renal impairment after the introduction of novel agents, Ann. Oncol. 25 (1) (2014) 195–200.

[8] G. Fouquet, B. Pegourie, M. Macro, M. Petillon, L. Karlin, D. Caillot, M. Roussel, B. Arnulf, C. Mathiot, G. Marit, et al., Safe and prolonged survival with long-term exposure to pomalidomide in relapsed/refractory myeloma, Ann. Oncol. 27 (5) (2016) 902–907.

[9] B. Ricci, M. van der Schaar, J. Yoon, E. Cenko, Z. Vasiljevic, M. Dorobantu, M. Zdravkovic, S. Kedev, O. Kalpak, D. Milicic, et al., Machine learning techniques for risk stratification of non-ST-elevation acute coronary syndrome: The role of diabetes and age, Circulation 136 (suppl_1) (2017) A15892.

[10] K. Ahuja, M. van der Schaar, Risk-stratify: Confident stratification of patients based on risk, 2018, arXiv preprint arXiv:1811.00753.

[11] B. Varghese, F. Chen, D. Hwang, S.L. Palmer, A.L. De Castro Abreu, O. Ukimura, M. Aron, M. Aron, I. Gill, V. Duddalwar, et al., Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images, in: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, 2020, pp. 1–10.

[12] E. Bria, G. De Manzoni, S. Beghelli, A. Tomezzoli, S. Barbi, C. Di Gregorio, M. Scardoni, E. Amato, M. Frizziero, I. Sperduti, et al., A clinical–biological risk stratification model for resected gastric cancer: prognostic impact of HER2, FHIT, and APC expression status, Ann. Oncol. 24 (3) (2013) 693–701.

[13] E. Hui, W. Li, B. Ma, W. Lam, K. Chan, F. Mo, Q. Ai, A. King, C. Wong, R. Guo, et al., Integrating postradiotherapy plasma epstein–barr virus DNA and TNM stage for risk stratification of nasopharyngeal carcinoma to adjuvant therapy, Ann. Oncol. 31 (6) (2020) 769–779.

[14] A. Farswan, A. Gupta, R. Gupta, S. Hazra, S. Khan, L. Kumar, A. Sharma, AI-supported modified risk staging for multiple myeloma cancer useful in real-world scenario, Transl. Oncol. 14 (9) (2021) 101157, http://dx.doi.org/10.1016/j.tranon.2021.101157.

[15] A. Farswan, A. Gupta, K. Sriram, A. Sharma, L. Kumar, R. Gupta, Does ethnicity matter in multiple myeloma risk prediction in the era of genomics and novel agents? Evidence from real-world data, Front. Oncol. 11 (2021) 4660, http://dx.doi.org/10.3389/fonc.2021.720932.

[16] S. Ailawadhi, I.T. Aldoss, D. Yang, P. Razavi, W. Cozen, T. Sher, A. Chanan-Khan, Outcome disparities in multiple myeloma: a SEER-based comparative analysis of ethnic subgroups, Br. J. Haematol. 158 (1) (2012) 91–98.

[17] A.J. Waxman, P.J. Mink, S.S. Devesa, W.F. Anderson, B.M. Weiss, S.Y. Kristinsson, K.A. McGlynn, O. Landgren, Racial disparities in incidence and outcome in multiple myeloma: a population-based study, Blood J. Am. Soc. Hematol. 116 (25) (2010) 5501–5506.

[18] L.J. Costa, I.K. Brill, J. Omel, K. Godby, S.K. Kumar, E.E. Brown, Recent trends in multiple myeloma incidence and survival by age, race, and ethnicity in the United States, Blood Adv. 1 (4) (2017) 282–287.

[19] B.A. Derman, J. Jasielec, S.S. Langerman, W. Zhang, A.J. Jakubowiak, B.C.-H. Chiu, Racial differences in treatment and outcomes in multiple myeloma: a multiple myeloma research foundation analysis, Blood Cancer J. 10 (8) (2020) 1–7.

[20] D.D. Alexander, P.J. Mink, H.-O. Adami, P. Cole, J.S. Mandel, M.M. Oken, D. Trichopoulos, Multiple myeloma: a review of the epidemiologic literature, Int. J. Cancer 120 (S12) (2007) 40–61.

[21] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D.E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (10) (2018) 749.

[22] A. Farswan, A. Gupta, TV-DCT: Method to impute gene expression data using DCT based sparsity and total variation denoising, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 1244–1248.

[23] A. Farswan, A. Gupta, R. Gupta, G. Kaur, Imputation of gene expression data in blood cancer and its significance in inferring biological pathways, Front. Oncol. (2020) 1442.

[24] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, 2016, CoRR abs/1606.09375.

[25] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, 2015, CoRR abs/1509.09292.

[26] W. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 30, Curran Associates, Inc., 2017.

[27] M. Niepert, M. Ahmed, K. Kutzkov, Learning convolutional neural networks for graphs, in: M.F. Balcan, K.Q. Weinberger (Eds.), Proceedings of the 33rd International Conference on Machine Learning, in: Proceedings of Machine Learning Research, 48, PMLR, New York, New York, USA, 2016, pp. 2014–2023.

[28] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner, G. Monfardini, The graph neural network model, IEEE Trans. Neural Netw. 20 (1) (2009) 61–80, http://dx.doi.org/10.1109/TNN.2008.2005605.

[29] K. Xu, W. Hu, J. Leskovec, S. Jegelka, How powerful are graph neural networks? 2018, CoRR abs/1810.00826.

[30] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2017, abs/1609.02907.

[31] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, Q. He, A comprehensive survey on transfer learning, 2019, CoRR abs/1911.02685.

[32] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.

[33] K.A. Frerichs, N.A. Nagy, P.L. Lindenbergh, P. Bosman, J. Marin Soto, M. Broekmans, R.W. Groen, M. Themeli, L. Nieuwenhuis, C. Stege, et al., CD38-targeting antibodies in multiple myeloma: mechanisms of action and clinical experience, Expert Rev. Clin. Immunol. 14 (3) (2018) 197–206.

[34] L. Kumar, N. Iqbal, A. Mookerjee, R.K. Verma, O.D. Sharma, A. Batra, R. Pramanik, R. Gupta, Complete response after autologous stem cell transplant in multiple myeloma, Cancer Med. 3 (4) (2014) 939–946.

[35] R. Gupta, L. Kumar, M. Dahiya, N. Mathur, P. Harish, A. Sharma, O.D. Sharma, V. Shekhar, Minimal residual disease evaluation in autologous stem cell transplantation recipients with multiple myeloma, Leukemia Lymphoma 58 (5) (2017) 1234–1237.

[36] R. Gupta, G. Kaur, L. Kumar, L. Rani, N. Mathur, A. Sharma, M. Dahiya, V. Shekhar, S. Khan, A. Mookerjee, et al., Nucleic acid based risk assessment and staging for clinical practice in multiple myeloma, Ann. Hematolo. 97 (12) (2018) 2447–2454.

[37] R. Sharma, A. Jain, A. Jandial, D. Lad, A. Khadwal, G. Prakash, R. Nada, R. Aggarwal, R. Ramachandran, N. Varma, et al., Lack of renal recovery predicts poor survival in patients of multiple myeloma with renal impairment, Clin. Lymphoma Myeloma Leukemia (2022).

[38] Y. Furukawa, J. Kikuchi, Molecular basis of clonal evolution in multiple myeloma, Int. J. Hematol. 111 (4) (2020) 496–511.