

Research Article

# Survival Analysis of Diabetic Patients Using Deephit with a Modified Sparsity Layer

James Rioba David\* , Herbert Imboga, Susan Mwelu

Department of Statistics and Actuarial Science, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

## Abstract

Deep Learning (DL) models for survival analysis have been employed to predict different long-term illnesses, such as diabetes, in the healthcare system. In these medical datasets, numerous features about a patient are recorded, which may limit the performance of these deep-learning models. The regularization approach is considered as a strategy to reduce network weights hence deep learning architectures have been proposed to handle the weights of the features in the soft feature selection technique. Deep networks such as Deephit have been employed in survival analysis to handle competing risks. In addition, the model can learn the survival distribution directly, however, the model's accuracy cannot be guaranteed when we have many irrelevant features recorded about a given patient. This research proposes to develop a novel model by exploiting the sparsity regulation technique on the fully connected layers. To achieve this, we modified DeepHit's architecture by adding a sparsity layer for the feature selection. Specifically, the Combined Group and Exclusive Sparsity Regularization were used for feature selection by exploiting sharing and competing relationships among network weights. The researcher trained the model using a diabetic dataset obtained from the Kaggle data repository and compared the efficacy of different models. The findings from the study revealed that the DeepHit-Combined Group Exclusive Sparsity (CGES) model achieved a more efficient network while at the same time improving its performance compared to other base networks with full weights. This research contributes valuable insights for regulators and medical practitioners in giving essential standards of care to diabetic patients to reduce complications, diabetic-related illnesses, and associated costs.

## Keywords

Artificial Neural Network (ANN), Combined Group and Exclusive Sparsity (CGES), Competing Risks (CR), Cumulative Incidence Function (CIF), Deep Learning (DL)

## 1. Introduction

Diabetes is a chronic illness that is caused by high glucose levels over a long period and can cause associated complications such as vision impairment or blindness, nerve damage, heart disease, stroke, and kidney disease if left undiagnosed [9]. There are three main categories of diabetes mellitus: Type I diabetes (insulin-dependent) is caused by insufficient pro-

duction of insulin; Type II diabetes affects how the body uses sugar (glucose) for energy, and gestational diabetes which occurs primarily during pregnancy [2, 10].

Diabetes poses a significant global health threat due to the increasing illness and death rates. In 2021, diabetes was responsible for 6.7 million fatalities and led to healthcare costs

\*Corresponding author: davidrioba04@gmail.com (James Rioba David)

**Received:** 10 March 2025; **Accepted:** 20 March 2025; **Published:** 31 March 2025



Copyright: © The Author(s), 2025. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

of at least \$966 billion, accounting for 9% of total adult spending [5]. Furthermore, the global diabetic population aged 20–79 reached 537 million in 2021 and is projected to rise to 643 million by 2030 and 783 million by 2045 [3, 14].

Kenya is also experiencing this emerging diabetic epidemic; the incidence of diabetes is estimated at 3.3% (821,500). If this trend is not checked, this figure is estimated to rise to 4.5% by 2030 [3].

A distinct feature of survival studies is that dependent variables can be censored. This study will adopt the right censoring for the patients who have only had the initial recovery, have dropped out of follow-up, have died, have been transferred to another hospital, and have not recovered at all during the study period [6].

In some cases, a subject may experience more than one event of interest known as Competing Risks (CR). For instance, a diabetic patient can die from different causes (e.g., heart disease, stroke, and kidney disease). The effectiveness of deep learning methods might be restricted in the presence of numerous irrelevant features [7]. This is particularly relevant in medical data sets, where a patient's information may encompass a wide array of characteristics. The regularization technique can be seen as a way to reduce network weights and deep learning architectures have been proposed to specifically address feature weights in soft feature selection [12].

In light of the advancements in artificial intelligence (AI) technology, there is a push to employ innovative methods like deep learning (DL) in anticipation of different long-term illnesses, such as diabetes [1]. The effectiveness of deep learning methods might be restricted in the presence of numerous irrelevant features [11]. This is particularly relevant in medical data sets, where a patient's information may encompass a wide array of characteristics. The regularization technique can be seen as a way to reduce network weights and deep learning architectures have been proposed to specifically address feature weights in soft feature selection [13].

The DeepHit technique has been employed to directly learn the distribution of survival times for survival analysis with competing risks. However, the model had only semi-adaptive layer sizes, with the number of hidden nodes being a fixed multiple of the input size [9]. The models, such as the Dynamic-DeepHit, showed improvement in discriminating individual risks of different forms of failure due to cystic fibrosis; this approach provided only static survival analysis. Deep networks require a large amount of memory and computation power to train. Further, the large number of parameters also means that the model is highly susceptible to over-fitting [9].

Exclusive sparsity has been used in a multi-task learning context. The main idea in the work is to enforce the model parameters for different tasks to compete for features, instead of sharing features on multi-task learning that leverages group sparsity [15]. Each pair of weights is given a different degree

of sharing and competition based on the similarity between the tasks given by a taxonomy. Thus, we propose to simply combine the group sparsity and the exclusive sparsity, which will result in a similar [16].

Our model adapted the previous deep learning survival model DeepHit's architecture and training procedure. The dataset used for this research was obtained from the Kaggle Dataset Repository. It contains a sample of reported and followed diabetes cases in the USA for 5 years (2018–2022) conducted by the Centers for Disease Control and Prevention (CDC). The combined group and exclusive sparsity regularization were used for feature selection by exploiting the sharing and competing relationships among different network weights, and the features selected were then used to fit our proposed model.

The findings of this study will play a significant role in providing valuable insights for regulators, and medical practitioners in giving essential standards of care to diabetic patients to reduce complications, diabetic-related illnesses, and their associated costs.

## 2. Deep Learning

Deep learning models mainly use mechanisms that rely on self-learning; therefore, they use ANNs (artificial neural networks), which usually copy the functioning of the human brain in terms of processing data [1].

### 2.1. Feature Selection in Machine Learning

Not all features in a dataset contribute to the performance of the model. Some of the features may be irrelevant or redundant. Feature selection in real-world machine learning tasks improves the accuracy and interpretability of the machine learning model [7].

### 2.2. Feature Selection in Deep Learning

In recent years, deep neural networks have made significant advancements in feature selection from medical datasets [11]. To extract features directly from the relationships between characteristics in diabetic data, we present a feature selection approach based on deep learning networks.

## 3. Materials and Methods

The study uses DeepHit and Combined Group and Exclusive Sparsity Regularization which was tested on a diabetic dataset obtained from the Kaggle data repository.

### 3.1. Survival Data

This study treated survival time as discrete and the time horizon as finite (e.g., no diabetic patients lived longer than a certain time, i.e., 100 years). The time set will be  $T = \{0, \dots,$

$T_{max}\}$ .

This study considered  $K \geq 1$  possible events of interest; we assumed that exactly one event occurs for each diabetic patient. The study considered right censoring for the diabetic patients who were lost to follow-up and died before recovery or before the study ended. The set of possible events is  $K = \{\emptyset, 1, 2, \dots, K\}$ , with  $\emptyset$  denoting right censoring.

This study considered a dataset  $D = \{(x^i, s^i, k^i)\}_{i=1}^N$  comprising survival data for  $N$  diabetic patients who have been followed up for a certain time, where  $x^i \in X$  is the vector of covariates.  $s^i = \min(T^i, C^i)$  is the time-to-event with  $T^i \in T$  and  $C^i \in T$  indicating the event and the censoring times respectively.  $k^i \in K$  is the event or censoring that occurred at the time  $T^i$ .  $T$  is either the time at which an event (e.g., death) occurred or the time at which the subject was censored.

### 3.2. Sparsity Regularization

Our study was inspired by linear models which achieve sparsity through  $L_1$ -regularization that proposed adding a sparse layer before the first hidden layer with one-to-one connections from the input layer [15]. We applied regularization to the weights of this new layer which means that our model encouraged lower weights and feature selection at the input level for the neural network.

#### 3.2.1. Group Sparsity Regularization for Deep Neural Networks

The main idea behind group sparsity regularization is to promote feature sharing among the network filters. In addition, group sparsity will reduce the complexity of the model by eliminating a neuron as a whole, which can thus help obtain practical speedups in deep neural networks [12, 1]. The group sparsity regularization is given below:

$$\Omega(W^l) = \sum_g \|W_g^l\|_2 = \sum_g \sqrt{\sum_i W_{g,i}^{(l)2}} \quad (1)$$

Where  $g \in G$  is a weight group,  $W_g^l$  is the weight matrix (or a vector) for group  $g$  that is defined on  $W^l$ , and  $W_{g,i}$  is a weight at index  $i$ , for group  $g$ .

#### 3.2.2. Exclusive Sparsity Regularization for Deep Neural Networks

Exclusive sparsity allows the model parameters for different tasks to compete for features instead of sharing features, as suggested by previous work on multi-task learning that leverages group lasso [16, 17]. The exclusive sparsity regularization is defined as:

$$\Omega(W^l) = \frac{1}{2} \sum_g \|W_g^l\|_1^2 = \frac{1}{2} (\sum_i \|W_{g,i}^l\|_1)^2 \quad (2)$$

where  $W_{g,i}^l$  is the instance of the sub-matrix (or the vector)  $W_g^l$ . This norm is often known to as the (1, 2) norm and is the 2-norm over the 1-norm group.

#### 3.2.3. Combined Group and Exclusive Sparsity Regularization

Inspired by the study conducted by [8], which highlighted that each pair of weights is given a different degree of competition and sharing based on the similarity between the tasks given by a taxonomy.

$$\Omega(W^l) = \sum_g \left\{ (1 - \mu_i) \|W_g^l\|_2 + \frac{\mu_i}{2} \|W_g^l\|_1^2 \right\} \quad (3)$$

where  $\Omega$  is the parameter that decides the entire regularization effect,  $W^l$  is the weight matrix for the  $l^{th}$  layer, and  $\mu_i$  is the parameter for balancing the sharing and competition term at each layer which is given by equation 4 below;

$$\mu_i = m + (1 - 2m) \frac{1}{L-1} \quad (4)$$

where  $L$  is the total number of all layers,  $l \in \{0, \dots, L-1\}$  is an index of each layer, and  $0 \leq m \leq 1$  is the lowest parameter value for the exclusive sparsity term.

### 3.3. Working Model

Our study proposes to combine the group sparsity and the exclusive sparsity, which will result in a similar effect (Combined Group and Exclusive Sparsity) for the shared sub-network as well as each of the cause-specific sub-networks. The model architecture is shown in Figure 1 below;

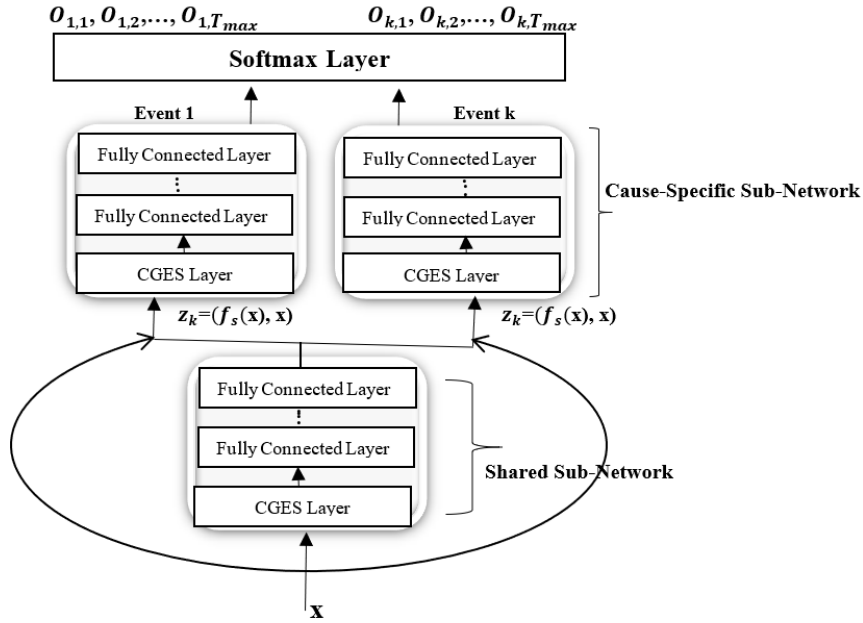


Figure 1. Network Architecture of our Model with K Competing Risks.

### 3.3.1. Shared Sub-network

The sub-network consists of the Combined Group and Exclusive Sparsity Regularization for feature selection. The  $L_s$  fully-connected layers make up the shared sub-network for  $k=1, \dots, K$ . A vector  $z_k$ , will capture the relationships shared by the K competing events, which is an output of the shared sub-network using the diabetes covariates  $x$  as inputs.

### 3.3.2. Cause-specific Sub-networks

The sub-network will estimate the joint distribution of the event time and competing events [4]. The  $k^{th}$  cause-specific sub-network for  $k = 1, \dots, K$  comprise of  $L_{c,k}$  fully-connected layers. Each cause-specific sub-network takes as inputs the pairs  $z_k = (f_s(x), x)$  and produces as output a vector  $F_{c,k}(z)$  which corresponds to the probability of a diabetic patient experiencing an event  $k$ .

### 3.3.3. Output Layer

To summarize the results of each cause-specific sub-network, DeepHit will use a soft-max layer. The output of the Softmax layer is a probability distribution  $y = [y_{1,1}, y_{1,2}, \dots, y_{1,T_{max}}, \dots, y_{K,1}, \dots, y_{K,T_{max}}]$ : an output element is the (estimated) probability  $\hat{P}(s, k|x)$  that the diabetic patient will experience the event  $k$  at time  $s$  given covariate  $x$ .

### 3.4. Cumulative Incidence Function (CIF)

To analyze cause-specific risk and make predictions, the study will use the cause-specific CIF. CIF is key to survival analysis under competing risks and it expresses the likelihood that a particular event  $k^* \in K$  occurs on or before time  $t^*$

conditional on covariates  $x^*$  [6]. The CIF for the event is given by:

$$F_{k^*}(t^*/x^*) = P(s \leq t^*, k = k^* | x = x^*) = \sum_{s^*=0}^{t^*} P(s = s^*, k = k^* | x = x^*) \quad (5)$$

Since the true CIF,  $F_{k^*}(t^*/x^*)$ , is unknown, the study will utilize the estimated  $F_{k^*}(t^*/x^*) = \sum_{m=0}^{s^*} O_{k,m}^*$ .

Similarly, the survival probability for the diabetic patient at the time  $t^*|X^*$  given as:

$$S(t^*/x^*) = P(T > t^* | x^*, t > t_j^*) = 1 - \sum_{k \neq \emptyset} F_{k^*}(t^*/x^*) \quad (6)$$

### 3.5. Training the DeepHit Model

To train the model, we will minimize the total loss function  $L_{Total}$  that is adjusted to include combined group and exclusive lasso regularizer term;

$$L_{Total} = L_1 + L_2 + \Omega(W^l) \quad (7)$$

where  $L_1$  is the log-likelihood of the joint distribution of the first hitting time and event while  $L_2$  incorporates a combination of cause-specific ranking loss functions, and  $\Omega(W^l)$  as defined in Equation 3. The loss functions  $L_1$  is defined as;

$$L_1 = \sum_{i=1}^N \left\{ \mathbb{1}(k^{(i)} \neq \emptyset) \cdot \log(y_{k^{(i)}}^{(i)}, s^{(i)}) + \mathbb{1}(k^{(i)} \neq \emptyset) \cdot \log[1 - \sum_{k=1}^K \hat{F}_k(s^{(i)} | x^{(i)})] \right\} \quad (8)$$

where  $\mathbb{1}(\cdot)$  is an indicator function. The first term will capture the information given by the uncensored diabetic patients, and the second term will capture the censoring bias. The  $L_2$  is

defined as;

$$L_2 = \sum_{k=1}^K \alpha_k \cdot \sum_{i \neq j} A_{k,i,j} \cdot \eta[\hat{F}_k(s^{(i)}|x^{(i)}), \hat{F}_k(s^{(j)}|x^{(j)})] \quad (9)$$

Where  $A_{k,i,j} \triangleq \mathbb{1}(k^{(i)}=k, s^{(i)} < s^{(j)})$  is an indicator function of pairs (i, j) for the event k,  $\alpha_k \geq 0$  is a hyperparameter chosen to trade off ranking losses of the  $k^{th}$  competing event and  $\eta(x, y)$  is a convex loss function. We will assume that the coefficients  $\alpha_k$  are all equal (i.e.  $\alpha_k = \alpha$  for  $k = 1, \dots, K$  and some  $\alpha$  to be chosen. The convex loss function can be expressed as;

$$\eta(x, y) = \exp \frac{-(x-y)}{\sigma} \quad (10)$$

### 3.6. Evaluation Metric

#### 3.6.1. Time-dependent Concordance Index ( $C^{td}$ -index)

The  $C^{td}$ -index will prevent the model from becoming overfitted as it requires the number of output nodes equivalent to  $|T|$ . The  $C^{td}$ -index is given by;

$$C^{td} = P\{\hat{F}_k(s^{(i)}|x^{(i)}) > \hat{F}_k(s^{(j)}|x^{(j)}) | s^{(i)} < s^{(j)}\} \quad (11)$$

The  $C^{td}$ -index for event k is determined by comparing pairs of diabetic patients in which one patient had experienced event k at a specific time but the other had not experienced any event of interest nor been censored at that point.

A perfect model that could correctly order every pair would have a concordance of 1, and a model that always orders incorrectly would have a concordance of 0.

#### 3.6.2. Time-Dependent Brier Score

The time-dependent Brier score is used to account for the unobserved true outcome status for individuals due to censoring. It measures the mean square error between the observed survival status and the predicted survival probability weighted by the inverse probability of censoring [12]. A lower Brier score of 0 indicates higher prediction accuracy.

$$BS(t) = \frac{1}{n} \sum_{k=1}^K [\hat{w}_i(t) \cdot \{1 - u^i\}^2 + \{1 - \hat{w}_i(t)\} \cdot \{0 - u^i\}^2] \quad (12)$$

where  $u^i = \hat{F}_k(s^{(i)}|x^{(i)})$  and  $\hat{w}_i(t)$  are the weights that account for censoring.

## 4. Results

Python was used for sample selection, descriptive analysis, categorization of data, and application of DeepHit and Com-

bined Group and Exclusive Sparsity Regularization. The research data was obtained from Kaggle consisting of a sample of reported and followed diabetes cases in the USA, for 5 years (2018–2022) conducted by the Centers for Disease Control and Prevention (CDC). The number of participants for each event is as follows:

*Total patients:* 39,977

*Event 1: Death from Diabetes:* 2,639 (6.60%)

*Event 2: Death from Stroke:* 3,509 (8.78%)

*Event 3: Death from Heart Disease/Attack:* 6,352 (15.89%)

*Censored Patients:* 27,477 (68.74%)

**Table 1.** Explanatory (Covariates) Variables.

No.	Explanatory Variables	No.	Explanatory Variables
1	Age of Patients	11	Types of Diabetes
2	Blood Glucose Level	12	Cholesterol Level
3	Sex	13	Marital Status
4	Past Family History	14	Employment
5	Regimen	15	Smoking
6	Blood Pressure	16	Any Healthcare Coverage
7	Physical Activity	17	General Health
8	Income	18	Level of education
9	Mental Health	19	Physical health
10	Body Mass Index (BMI)		

Normally, a diabetes data set has some missing data. Before fitting our model, it was inevitable to carry out data pre-processing for missing data to upgrade the model's performance in forecasting.

In this research, we carried out data preprocessing by calculating missing values. Taking into consideration the kind of data that is missing in the data set that is applied in the analysis, mean values are used to stand in for the values that were missing in the case of numeric variables. All features were furthermore normalized to mean 0 and variance 1 before training the models.

### 4.1. Feature Importance

The model's architecture focuses on learning weights for input characteristics across shared and cause-specific sub-networks. The regularized weights ensure that large weights are only used when necessary to increase predictive value.

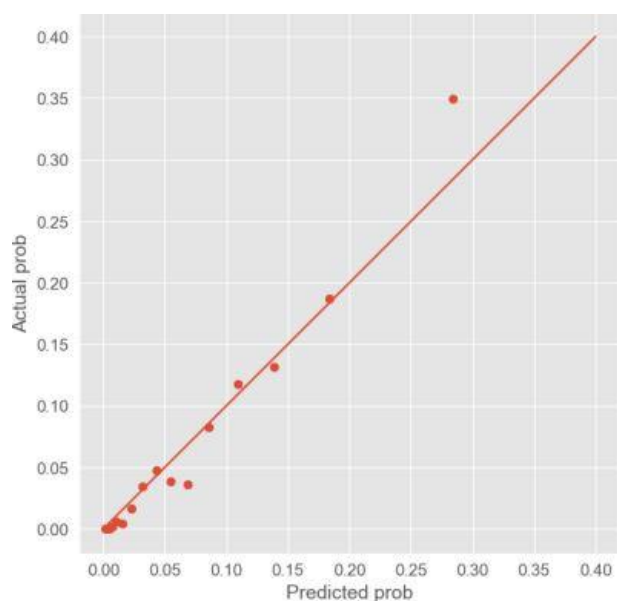


**Table 2.** Top 10 Most Influential Features Determined by the Weight's Absolute Value (WAV).

Causes of Death						
Rank	Diabetes		Stroke		Heart Disease	
	Feature	Importance (WAV)	Feature	Importance (WAV)	Feature	Importance (WAV)
1	Age	0.151461	Age	0.215366	Age	0.244801
2	BMI	0.140506	Sex	0.147312	Diabetes Type	0.167733
3	Sex	0.127359	BMI	0.091544	Sex	0.068167
4	Diabetes Type	0.097647	Family History	0.077154	BMI	0.058409
5	Family History	0.075131	Smoker	0.064789	Family History	0.054690
6	Blood Glucose	0.062676	Blood Pressure	0.042899	Blood Pressure	0.037006
7	Blood Pressure	0.053332	General Health	0.032045	Smoker	0.032934
8	Education	0.046267	Diabetes Type	0.020033	Education	0.027868
9	Income	0.028818	Marital Status	0.019821	General Health	0.021602
10	General Health	0.017610	Education	0.014161	Income	0.018450

Table 2 above illustrates the most influential covariates for death from Diabetes, Stroke, and death from heart disease respectively using Combined Group and Exclusive Sparsity feature selection. After setting  $\mu_1 = 0.8$  [15], the first 10 features were selected and they were used to fit our proposed model. Eliminating less pertinent features enhances the interpretability of the model features and also improves the overall accuracy of the models.

## 4.2. Calibration Plots

**Figure 2.** Calibration Plot for Our Model.

The Calibration plot of our model for the selected feature is presented in Figure 2. The calibration plot indicates a line of a perfectly calibrated model at the closest; in particular, the diabetic patient gets a predicted probability of experiencing an event of interest around  $\approx 20\%$ , which coincides with the actual event status.

## 4.3. Survival and Probability Curves

We utilized survival data to conduct survival analysis by predicting mortality using deep learning algorithms. The cause-specific data is continually updated at each time interval  $T$ , ranging from 0 to  $T_{max}$ , to include new patient datasets of those who are censored, and deceased from diabetes, stroke, and heart disease up to the  $T^{th}$  time interval. With our dataset  $D = \{(x^i, s^i, k^i)\}_{i=1}^N$ , we managed to calculate the risk score of a patient at every time interval.

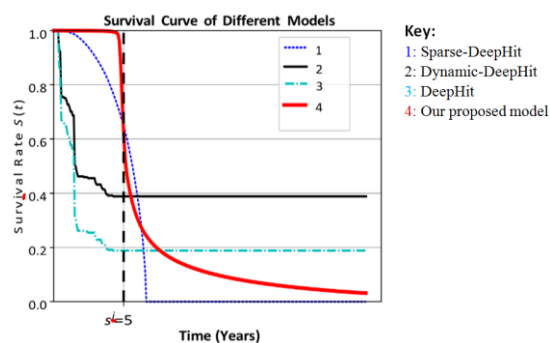
**Figure 3.** Survival rate  $S(t|x^i)$  estimation over different models. The vertical dotted line is the true event time  $s^i$  of this sample.

Figure 3 above illustrates the estimated survival rate curve over time for the test sample  $(x^i, s^i, k^i)$ . Our model accurately placed the highest survival scores on the true event time  $t$  with survival scores. The sparse-DeepHit model on the other

achieved a fair prediction on the survival on the true event time  $t$  but Deephit achieved the lowest prediction score on survival on the true event time  $t$ .

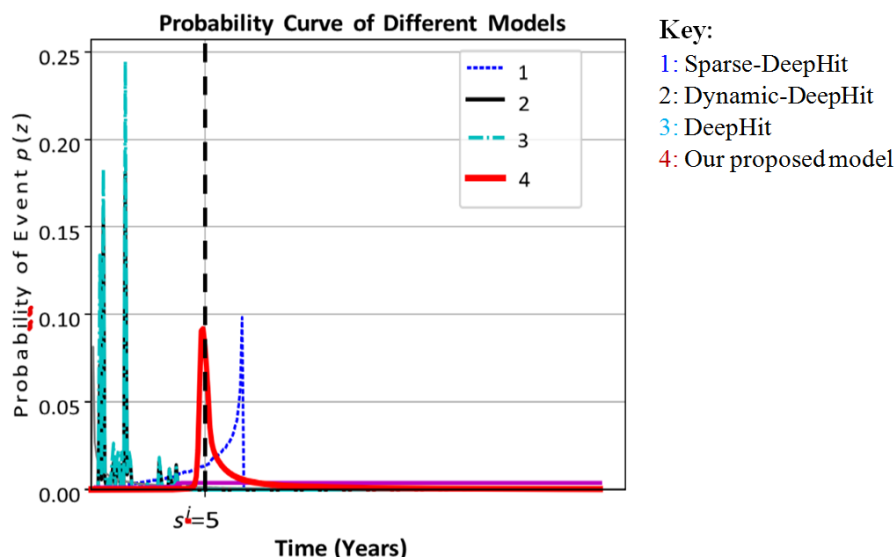


Figure 4. Event Time Probability Prediction over Different Models. The vertical dotted line is the true event time  $t$  of this sample.

Figure 4 above illustrates the estimated forecasted event time probability for the test sample  $(x^i, s^i, k^i)$ . Our model accurately placed the highest probability on the true event time  $t$  with high probability scores.

#### 4.4. Evaluation of the Models' Performance

Table 3. Time-Dependent Concordance Index Performance of Sparse-DeepHit, Dynamic-Deephit, Deep Hit, and our Proposed Model. SD denotes the Standard Deviation.

Event/Model	Diabetes		Stroke		Heart Disease	
	Mean	SD	Mean	SD	Mean	SD
Sparse-DeepHit	0.760	0.027	0.773	0.012	0.780	0.011
Dynamic-DeepHit	0.740	0.031	0.765	0.014	0.779	0.013
DeepHit	0.684	0.038	0.691	0.026	0.752	0.020
Proposed Model	0.770	0.022	0.790	0.010	0.799	0.088

The results in Table 3 show that our model achieved significantly better performance than the other three models across all the events. Specifically, the best prediction for diabetes (heart disease) attained a higher index of 0.799 compared to 0.780, 0.779, and 0.752 for Sparse-DeepHit, Dynamic-DeepHit, and DeepHit, respectively.

#### 4.5. Time-Dependent Brier Score

Table 4 indicates that our model achieves a better performance of 0.068, 0.065, and 0.063 across all events comparable to the three models. Specifically, our model achieved a better Brier score that is close to zero across all the events; 0.059 for death from heart attack compared to 0.063, 0.066,

and 0.075 for Sparse-DeepHit, Dynamic-DeepHit, and DeepHit, respectively.

**Table 4.** Time-Dependent Brier Score Performance of Sparse-DeepHit, Dynamic-Deephit, DeepHit, and our Proposed Model. MSE denotes Mean Squared Error.

Event/Model	Diabetes	Stroke	Heart Disease
	MSE	MSE	MSE
Sparse-DeepHit	0.071	0.065	0.063
Dynamic-DeepHit	0.074	0.069	0.066
DeepHit	0.078	0.076	0.075
Proposed Model	0.068	0.064	0.059

## 5. Discussion

The main objective of this research was to explore the use of feature selection techniques on DeepHit architecture. To do this, we used Combined Group and Exclusive Sparsity Regularization for feature selection and we trained our model using data obtained from the Kaggle data repository. The combined Group and Exclusive Sparsity Regularization technique was used because it is capable of leveraging the sharing and competing relationships among different network weights. After setting  $\mu_i = 0.8$  [15], the first 10 features were selected and they were used to fit our proposed model. The study's findings demonstrated that using the model's design, the built model attained the highest time-dependent index of 0.799 and the lowest time-dependent Brier score of 0.059.

## 6. Conclusions

We made improvements to the previous deep learning survival model, DeepHit, by adapting its architecture and training procedure. The model was trained on diabetic data obtained from the Kaggle online data repository. The Combined Group and Exclusive Sparsity Regularization were used for feature selection by exploiting the sharing and competing relationships among different network weights. The findings from the study revealed that the DeepHit-CGES model achieved more efficient networks while improving its performance compared to other base networks with full weights. Whilst our model applies to survival analysis, we also see the potential for other layers in the deep neural network such as convolutional networks.

## Abbreviations

AI Artificial Intelligence

ANN Artificial Neural Network  
 CDC Center for Disease Control and Prevention  
 CGES Combined Group and Exclusive Sparsity  
 CIF Cumulative Incidence Function  
 CR Competing Risks  
 DL Deep Learning

## Author Contributions

**James Rioba David:** Conceptualization, Data curation, Formal Analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing

**Herbert Imboga:** Supervision

**Susan Mwelu:** Supervision

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

- [1] Alvarez, J. M., & Salzmann, M. (2016). Learning the number of neurons in deep networks. *Advances in neural information processing systems*, 29. <https://doi.org/10.48550/arXiv.1611.06321>
- [2] American Diabetes Association. (2021). Classification and diagnosis of diabetes: standards of medical care in diabetes—2021. *Diabetes care*, 44 (Supplement\_1), S15-S33. <https://doi.org/10.2337/dc21-S002>
- [3] Atlas, D. (2021). International diabetes federation. IDF Diabetes Atlas, 10th ed. Brussels, Belgium: *International Diabetes Federation*, 33(2).
- [4] Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, 160-167. <https://doi.org/10.1145/1390156.139017>
- [5] Federation ID. (2021) Diabetes atlas. 10th Edition. Brussels, Belgium.
- [6] Fine, J. P., & Gray, R. J. (1999). A Proportional Hazards model for the sub-distribution of a competing risk. *Journal of the American Statistical Association*, 94(446), 496–509. <https://doi.org/10.2307/2670170>
- [7] Kim, S., & Xing, E. P. (2010). Tree-guided group lasso for multi-task regression with structured sparsity ICML. Google Scholar Google Scholar Digital Library Digital Library, 6(3), 1095-1117. <http://dx.doi.org/10.1214/12-AOAS549>
- [8] Lee, C., Yoon, J., & Van Der Schaar, M. (2019). Dynamic-Deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, 67(1), 122-133. <https://doi.org/10.1109/TBME.2019.2909027>



- [9] Lee, C., Zame, W., Yoon, J., & Van Der Schaar, M. (2018, April). Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1). <https://doi.org/10.1609/aaai.v32i1.11842>
- [10] Ormazabal, V. et al. (2018). Association between insulin resistance and the development of cardiovascular disease. *Cardiovascular Diabetology*, 17, 122. <https://doi.org/10.1186/s12933-018-0762-4>
- [11] Rietschel, C. (2018). Automated feature selection for survival analysis with deep learning (Doctoral dissertation, University of Oxford). <https://doi.org/10.1007/s10462-023-10681-3>
- [12] Schoop, R., Beyersmann, J., Schumacher, M., & Binder, H. (2011). Quantifying the predictive accuracy of time - to - event models in the presence of competing risks. *Biometrical Journal*, 53(1), 88-112. <https://doi.org/10.1002/bimj.201000073>
- [13] Wen, W., Wu, C., Wang, Y., Chen, Y., & Li, H. (2016). Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29. <https://doi.org/10.48550/arXiv.1608.03665>
- [14] World Health Organization. (2021). Diabetes Fact Sheet. <https://www.who.int/news-room/fact-sheets/detail/diabetes>
- [15] Yoon, J., & Hwang, S. J. (2017). Combined Group and Exclusive Sparsity for Deep Neural Networks. *International Conference on Machine Learning*, 7, 81-89. <https://doi.org/10.1016/j.neucom.2017.02.029>
- [16] Yu, C-N., Greiner, R., Lin, H-C., & Baracos, V. (2011). Learning patient-specific cancer survival distributions as a sequence of dependent regressors. In *Advances in Neural Information Processing Systems* 24. <https://doi.org/10.3389/fonc.2022.967758>
- [17] Zhou, Y., Jin, R., & Hoi, S. C. H. (2010, March). Exclusive lasso for multi-task feature selection. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 988-995). JMLR Workshop and Conference Proceedings. [https://ink.library.smu.edu.sg/sis\\_research/2317](https://ink.library.smu.edu.sg/sis_research/2317)