Survival Analysis Based Qos Recommendation for Bus Transportation Using Deep Learning

Himabindu N[1], Samyama Gunjal G H[2], Nagendrababu N C[3]

[1,3]Student, University of Visvesvaraya College of Engineering
[2]Associate Professor, University of Visvesvaraya College of Engineering

Abstract:
Public transportation play an important role in metropolitan cities.Bus services are an important part of public transportation systems, which are frequently chosen because they are convenient, economical, environmentally beneficial, and less taxing on the infras- tructure. The temporal patterns of bus transportation incidents are examined in this study using survival analysis, with a focus on the amount of time until certain events like delays,malfunctions, or safe arrivals at destinations. High-dimensional data is being generated at an increasing rate due to technological advancements. As a result, it is quite difficult to ana-lyze such a dataset adequately. Algorithms for machine learning (ML) have become popular tools for modeling complex and nonlinear interactions in a variety of real-world applica- tions, including as the analysis of high-dimensional survival data. Multilayer Deep Neural Network (DNN) models have achieved impressive results recently. Thus, using Keras and TensorFlow, a Cox-based DNN prediction survival model (DNNSurv model) was created. Its findings, however, were limited to survival datasets with large sample sizes or high di- mensionality. Using high dimensional survival data, the proposed work will evaluate the DNNSurv model's prediction ability and compare it to a well-known machine learning sur- vival prediction model (random survival forest). The Proposed Work also offers the best settings for a number of hyperparameters, including tuning parameter selection, for this reason. The suggested approach showed through data analysis that, when compared to the ML model, the DNNSurv model performed better overall in terms of the two primary sur- vival prediction evaluation measures (i.e., the concordance index and the time-dependent AUC).

Keywords: DNNSurv, Deep Neural Networks,Random Survival Forest, Survival Analysis.

1. Introduction

The quality of bus services greatly impacts urban mobility. Hence, understanding how timely and reliable these buses are is essential. To optimize service quality, it is important to understand the time-to-event dynamics of bus transportation events, such as successful arrivals, breakdowns, and delays. This study utilized survival analysis techniques to examine such patterns within a comprehensive bus transportation dataset. Various fields, including healthcare, finance, and engineering, rely on survival analysis as a statistical approach to explore time-to-event data. It is a significant method that emphasizes comprehending the length until a specific occasion takes place, therefore predicting its progression. Identifying factors that may influence the timing of an event is critical in survival analysis, specifically in estimating the survival function. The survival function represents the probability that an event will not happen by a certain time and is a helpful tool in understanding the distribution of event times. The dataset combines a variety of sources, including bus schedules, GPS data, and historical performance records. Rigorous preprocessing ensures that the dataset is ready for Survival analysis, and handles issues such as missing values. Survival analysis models, such as the Kaplan-Meier estimator and Cox proportional hazards model, have been widely used in various fields to analyze event occurrences and predict their outcomes. However, the scope of this study goes beyond traditional methodologies and incorporates the Random Forest method, which is renowned for its capacity to capture intricate relationships and handle diverse features.

By employing the Random Forest model, this study aims to discern patterns in bus transportation events and establish a robust foundation for event prediction. Because of censoring, the exploration of survival data is gener- ally far more intricate than reg- ular statistical analysis. Numerous statistical approaches have been devised for the scrutiny of survival, employing often non-parametric or semi-parametric statistical techniques. Public transporta- tion, like bus transportation, has a crucial function in urban mobility, bestowing an effective and cost-effective means of conveyance for countless individuals daily. Nonetheless, guaranteeing top-notch assistance for com- muters can prove to be quite a challenge, given diverse factors such as traffic congestion, weather circumstances, and unforeseen incidents. In the realm of Bus Transportation, guaranteeing a dependable and punctual service is crucial to enhance passenger contentment and stimulate heightened ridership. To accomplish this, a groundbreak- ing strategy that melds survival analysis and deep learning can be utilized to provide individualized Quality of Service (QoS) suggestions for bus passengers. The Conceptualized Undertaking examined the HD case, where p is exceptionally voluminous. Lately, machine learning (ML) algorithms have been extensively implemented to model non-linear and intricate interplays and to enhance prognostication, in a multitude of pragmatic domains. The application of neural net- work algorithms in survival analysis, particularly in handling incomplete data for HD survival data, has been notable. Over time, the multilayer deep neural network (DNN) model has demonstrated significant advancements, particularly in addressing complex and high-dimensional cases with complete data The goal is to use the simplicity and presentation of deep learning models to capture physical patterns in QoS profiles. Also deep neural network (DNN) survival analysis model. Deep learning has the advantage of capturing complex temporal dependencies and nonlinear relationships in information, especially in the context of survival analysis. The DNN survival analysis model aims to reveal the complex patterns present in bus traffic situations and increase the understanding of the factors affecting the reliability of services. The combination of random forest methods and DNN survival analysis represents a two-pronged approach that aims to ensure that the classifier is interpreted with the ability of deep learning to handle complexity. Based on this dual model concept, this researchaims to analyze the nature of the bus and provide an understanding beyond survival analysis. The results of this study are expected to not only improve the prediction accuracy of bus status, but also provide a better understand- ing to improve the overall performance and reliability of urban bus services. Moreover, realizing the potential of deep learning to capture changes and non-linear relationships in the body's body, the deep neural network network (DNN) survival analysis model was introduced. Neural network architecture is designed to delve into the rich complexity inherent in traffic data and facilitate our ability to analyze the subtle interactions between factors af- fecting service. Urban transportation relies on the efficiency and reliability of transportation buses, making quality of service (QoS) an important factor in the overall operation. It requires new analytical methods regarding time precision and dynamic aspects of QoS metrics such as bus arrival time, waiting time and services.

This study presents a method to combine survival analysis using correlation coefficient (C-index) and area under the curve (AUC) indices, especially in the context of the bus. Survival analysis is widely used in medical research, where it is suitable for modeling time-to-event QoS metrics on the transport bus. The C-index is a measure of the agreement between predicted and observed survival and becomes an important index. The C-index indicates the ability of the model to predict survival time, making it a good indicator to evaluate the accuracy of the survival model in the context of bus transportation. In addition to the C-index, the distinctiveness of the survival model was evaluated using the area under the curve (AUC). AUC is often related to the receiver operating characteristic (ROC) curve, which provides a general measure of the model's ability to discriminate between different survival rates. It shows that the model can distinguish between different QoS levels, helping to identify important factors affecting the reliability of the service. The unique contribution of this study is the combination of normal survival analysis, random forest, and DNN models. By leveraging the power of each method, we not only predict events more accurately but also seek to better understand the complexity of urban bus traffic. Integration of survival analysis, C-index and AUC into bus data not only leads to a detailed understanding of QoS dynamics, but also leads to the development of traffic patterns estimators useful in capturing physical dependency and dynamic patterns. Using these metrics, the goal is not only to evaluate the reliability of on-time bus service but also to provide tools for future forecasting and decision-making. The integration of survival analysis, random forest and DNN models into the bus dataset aims not only to provide a better understanding of service reliability, but also to provide benchmarks of the prediction of advanced models. C-index and AUC work as important parameters and provide a strong basis for evaluating the performance of each model and guide the selection of the best transport system.QoS prediction methods. This study has important implications for education and business, not only to provide a theoretical understanding of the power of good service, but also to provide practical tools to increase city confidence in transport authorities. bus service safety and efficiency.

2. LITERATURE REVIEW

Clark TG, Bradburn MJ. et al. [1] High-dimensional data, where the number of predictors exceeds the number of observations, present challenges for classical survival regression models. These models are often either impractical to fit or prone to overfitting, resulting in low predictability. While deep learning has shown promise in survival analysis for censored data, existing approaches typically utilize only a single hidden layer, limiting their effectiveness.To address this limitation, we propose the DNNSurv model, which employs deep neural networks with multiple hidden layers.

The paper [2] Anticipating urban dispersal events is crucial for effective urban planning and management. This study introduces a novel Two-Stage Framework utilizing deep survival analysis on mobility data to enhance the predictive capabilities of such events.The outcomes of this research have the potential to rev- olutionize urban planning by providing a data-driven approach to anticipate and mitigate the impact of dispersal events. The Two-Stage Framework offers a versatile and adaptable solution, contributing to the resilience and sustainability of urban environments.

The paper [3]Survival analysis is crucial in various fields, such as healthcare and finance, where predicting time-to-event outcomes is of paramount importance.The research begins by delving into the fundamentals of survival analysis and the challenges associated with modeling complex temporal relationships. Deep neural networks are introduced as a powerful tool to capture intricate patterns within survival data. To address the interpretability concern, the study incorporates the concept of pseudo values, which are surrogate representations of the true survival times.

The work [4] The application of this approach extends to various domains, such as genomics, where the number of features often exceeds the number of observations. By effectively combining the strengths of the ELM and the Cox model, this research aims to provide a robust solution for extracting meaningful insights from high-dimensional survival data, contributing to advancements in precision medicine and other fields.
The paper [5] explores the application of deep neural networks in survival analysis within the context ofa multi-task framework. Survival analysis, crucial in various fields like healthcare and finance, focuses on predicting time-to-event outcomes. The research integrates deep learning techniques into a multi-task setting to enhance the model's versatility and performance.The Work [6] introduces a cutting-edge approach to survival analysis tailored for high-dimensional survival data, leveraging the power of deep learning. In fields like genomics and biomedical research, where datasets are characterized by a large number of features, traditional survival models may face challenges. The study proposes a novel framework that employs deep learning techniques to effectively handle high-dimensionalsurvival data.
The Paper [7] Random Survival Forests, an extension of traditional Random Forests, are introduced as an ensemble learning method capable of handling censored observations and capturing non-linear relationships in survival data. the construction of the RSF model, and an exploration of its advantages, including resilienceto over-fitting and the ability to handle high-dimensional covariates.
[8] Applications of this approach extend to fields such as genomics, where the number of genetic markers far exceeds the sample size. By seamlessly integrating deep learning into survival analysis, research seeks to provide a robust and accurate tool for extracting meaningful insights from high-dimensional datasets, contributing to advancements in personalized medicine and other related domains.
[9] explores the application of neural networks in both continuous and discrete-time survival prediction, offering a versatile approach to modeling time-to-event outcomes. Survival prediction plays a crucial role in diverse fields such as healthcare and finance, and this study introduces a framework that seamlessly inte- grates neural networks to address the challenges associated with both continuous and discrete-time settings.
[10] Classical survival analysis methods often assume linear relationships between predictors and the hazard function. Deep learning models, on the other hand, can capture non-linear patterns in the data, allowing formore flexible and accurate representations.
The overview begins by emphasizing the significance of survival prediction and the complexities involved in handling time-to-event data. Neural networks are introduced as a powerful tool capable of capturing intricate patterns and dependencies within the data, providing a flexible solution for survival analysis.

Literature Survey				
SL.no	Tittle	Author	Advantage	Disadvantage
1	Survival analysis partI: Basic concepts and first analyses.	Clark TG, Brad-burn MJ, Love SB, Altman DG	It provides a more compre- hensive understanding of the data by considering the time aspect, which is crucial in many real-world scenarios. Allows for the	The Cox proportional haz- ards model assumes that the hazard rate remains constant over time. Violation of this assumption affects the valid-ity of the results.

			inclusion of covariates to assess their im-pact on the survival outcome	
2	Deep survival mod-elling for shared mo-bility .	Bojan Kostic, Mathilde Pryds Loft, Filipe Rodrigues	Deep survival models can handle complex temporal spatiotemporal patterns in shared mobility data.By leveraging deep learning techniques, these models often achieve better predic- tive accuracy compared to traditional methods	Deep survival models can lead to overfitting, espe-cially when dealing with small or noisy data. Over- fitting can damage new data and reduce model confi-dence.
3	Predicting Urban Dispersal Events: A Two-Stage Framework through Deep Survival Analysis on Mobility Data	Amin Vahedian, Xun Zhou, Ling Tong, W. Nick Street	Deep survival analysis mod- els can effectively predict urban dispersal events, such as mass migrations or crowd dispersals, with high ac-curacy. The framework al- lows for capturing complex spatiotemporal patterns, im- proving the precision of event predictions.	Finding the optimal hyper- parameters for deep learn- ing models can be time- consuming and computa- tionally expensive. Exten- sive tuning may be neces-sary to achieve the best per-formance.
4	Deep Neural Networks for Survival Analysis Using Pseudo Values	Zhao L, Feng D	Corporates pseudo values for efficient model train-ing.Provides flexibility in handling various types of predictors, including high- dimensional genomic data.	Requires large and diverse datasets for training deep neural networks effec-tively.Interpretability and transparency of deep neural networks.
5	Deep Neural Networks for Survival Analysis Based on a Multi-TaskFramework	Stephane Fotso	Utilizes a multi-task frame- work for survival analy- sis. Incorporates deep neural networks for learning com- plex relationships. Consid- ers multiple related survival tasks simultaneously.	Potential overfitting with small sample sizes or imbal- anced datasets.Challenges in validation and repro-ducibility of multi-task deep neural network models in survival analysis tasks.
6	Extreme learning ma- chine Cox	Li G, Wang H	Suitable for high-dimensional survival	Lack of theoretical guar-antees for the ELM algo-

	model for high-dimensional sur- vival analysis		anal- ysis, where the number of predictors is large.Efficient training and prediction process due to the extreme learning machine (ELM) algorithm.Does not require manual tuning of hyperpa- rameters, making it easier touse.	rithm, which may affect the model's reliability. Depen- dency on the quality and rep- resentativeness of the train- ing dataset.Limited avail- ability of software packages and resources specifically designed for the ELM Cox model, which may limit its implementation and adop- tion.

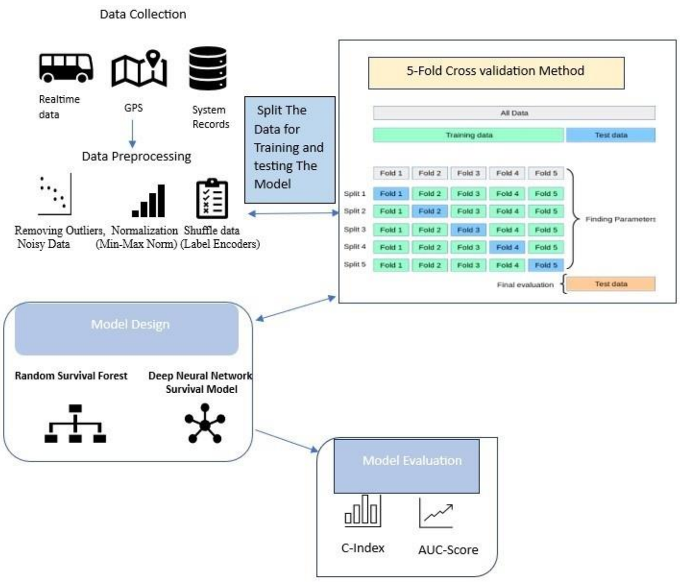

Fig. 1. Architecture of Proposed System

3. PROPOSED SYSTEM

Survival analysis is a method used to determine the time until an event of interest occurs. In the context of Trans- port Quality of Service (QoS) recommendations, survival analysis can be used to model time. This analysis can be used to make personalized recommendations to travelers. The working concept is to use DNNSURV deep learning based on the KERAS algorithm to train the dimensions and then compare its performance with existing learning machines such as Random Survival Forests. The main point of the proposed work The aim is to use existing ones to evaluate survival time analysis of deep learning and analyze C-index and time-dependent Auc curves, etc. Al- gorithms to evaluate Qos recommendations by measuring. We only use traffic data in our planning efforts. Gather comprehensive bus transportation data, including schedules, real-time GPS information, historical performance records, and any relevant contextual data such as weather conditions and traffic patterns.

The architecture can vary depending on the specific approach and techniques used, but here is a generalized overview of the key components

Data Collection: In the proposed work mainly used New York City (NYC) bus dataset. The Bus

Transporta- tion Data Set. Gather comprehensive bus transportation data, including schedules, real-time GPS informa- tion, historical performance records, and any relevant contextual data such as weather conditions and traffic patterns. In which The Dataset is of 1.3 GB Which can't be shown in the Notepad in which it can be Rep- resented By Graph As Shown Below Figure. When Loaded dataset can seen traffic information, bus arrival, schedules and expected time etc. By Representing The NYC Dataset by graph where the x-axis represents Place Name and y-axis represents count of that places visited by buses. The data set consists 5000 rows and 17 Columns. The Figure 2 is the dataset where can be see x-axis represents place names and y-axis is how many times the bus visited those places.

Data Pre-Processing: Involves preparing and transforming the raw data into a suitable format that can be ef- fectively utilized by the models. Some data prepossessing steps for survival analysis with high-dimensional data are Done Like remove missing values, Feature Extraction Like Normalization of Data set, Data Clean-ing and shuffle data set values. It identifies and select relevant attributes or features from the dataset, such as Passenger Gender, Passenger age, departure times, arrival times, distance, number of stops, etc. Choose features that are meaningful for specific analysis or prediction task like Arrival Time, Expected Time, Places Names, Distance, Number of Bus stops etc. In Normalization Scale numerical features to a consistent range, often between 0 and 1, to prevent certain features from dominating others. Normalization aids in achieving uniformity and enhances the performance of machine learning models, especially those sensitive to varying scales. Randomly reorder the dataset to remove any inherent sequence. Shuffling is important to prevent the model from learning patterns based on the order of data entries. This ensures unbiased training and improves the generalization ability of the model.

Model Selection: It involves choosing the appropriate model architecture and configuration that can ef-

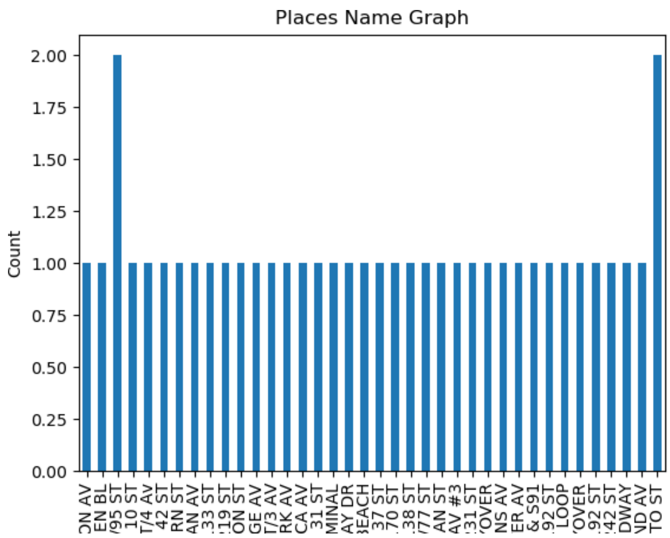

Fig. 2. Data set

fectively capture the patterns and relationships in the data and make accurate predictions for survival out- comes.In Proposed Work existing Machine Learning algorithms such Random Survival Forest and Deep Learning Model Such Has DNNSURV Model

Model Training: Model training in survival analysis using deep neural networks involves the process of optimizing the model's parameters to minimize the prediction error and improve its ability to estimate survival times or event probabilities accurately. To optimize the model parameters 5 CV-Fold method is used Cross-validation is a technique used to assess the performance and generalization ability of a predictive model. In the context of survival analysis with a bus transportation dataset, 5-fold cross-validation involves dividing the dataset into five approximately equal-sized subsets or folds. where the Splitting of the dataset To Test and Train the models. 5-fold cross-validation is a widely used method in survival analysis for evaluating the robustness and generalization performance of predictive models applied to bus transportation datasets. It provides a balance between computational efficiency and reliable model assessment, helping practitioners make informed decisions about model selection and hyperparameter tuning.

Model Interpretation: The combination of C-index and time-dependent AUC provides a comprehensive evaluation of a survival analysis model applied to a bus transportation dataset. These metrics offer insights into the model's ability to discriminate between different survival times and its performance across various temporal stages, contributing to informed decision-making in transportation planning and management. Given the time-dependent nature of survival analysis in the transportation domain, it's crucial to consider how well the model predicts events over time. Both C-index and time-dependent AUC should account for censored data, common in survival analysis when events have not occurred by the end of the observation period. Time-dependent AUC is beneficial for assessing how well the survival model predicts events, such as delays or breakdowns, at different time intervals. This can be particularly relevant in the context of transportation planning and management.

3.1. *Random Survival Forest*

RSF Model excels in effectively managing intricate and non-linear connections within bus transportation data. It also adeptly incorporates censored data, a common occurrence in survival analysis datasets. Moreover, the model ensures interpretability through feature importance measures [11]. The RSF model has consistently showcased a favorable average C-index, signifying its strong discriminatory power in accurately predicting survival times. Additionally, the model's average time-dependent AUC score further highlights its proficiency in ranking survival times across various intervals. This method are explained in algorithmic form. Due to the use of a binary tree, node splitting, and prediction are mentioned. The cumulative hazard function for the out-of-bag (OOB) set is described. The variable importance in prediction (VIMP) and its error are also specified here. Random Survival (RSF) was introduced to extend the RF to the censored survival data set. The implementation of RSF is as follows. The same general principles of RF:

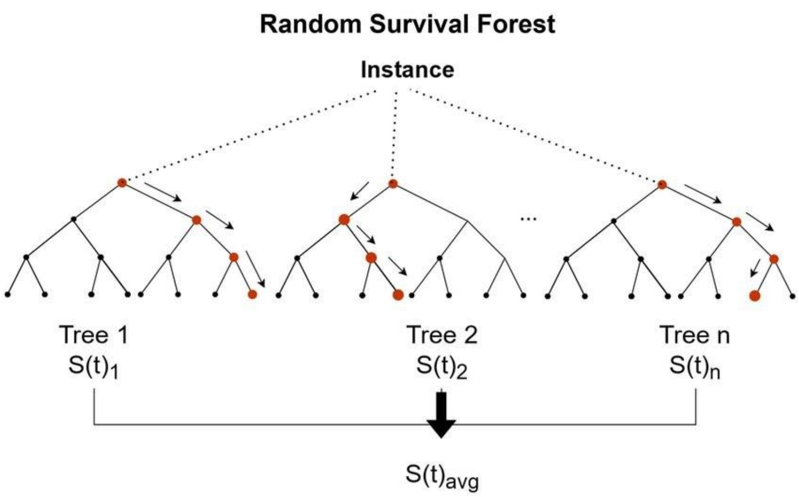

Fig. 3. Random Survival forest structure

1. The survival tree is grown using bootstrap data.
2. Random feature selection is used when splitting tree nodes.
3. Trees are grown deeply,
4. Forest group survival is calculated from the average of port number statistics (TNS).

Censorship is a unique feature of survival data and complicates certain aspects of implementation. We compared RSF with RF in terms of regression and classification. For right-censored survival data, the observed data are (T,δ). where T is time and δ is the censoring indicator. The observed time T is defined as the minimum value real time. survival event time (probably unobserved) to censored and true (probably unobserved) firm time; So T=min(To,Co) and the actual time of the event may not be observed. Censor index is defined

$$\delta = I\{To \leqslant Co\} \quad (1)$$

If δ =1, the event occurs (i.e. death occurs) and the actual event time T=To is observed. Otherwise, if δ =0, the observations are censored and only the censoring time T=Co is observed. The subject survived until Co, but not when the subject actually died. Henceforth it is called data. Because of (T1,X1,δ 1),. . . ,(Tn,Xn,δ n) where Xi is the feature vector (auxiliary variable) of individual i, Ti and δi. Observation time and censoring index for i. RSF trees are grown using resampling like RF trees. When When growing an RSF tree, you need to monitor the real-time events that are censored. Specifically, partitioning The rules for tree growth should be specifically considered for pruning. So the goal is to divide the tree nodes into: Left and right subsidiaries behave differently in the event (survival) history. The default division rules used by The statistic package is logrank test with splitrule=”logrank”. The log-rank test is traditionally used Although intended for two-sample tests using survival data, it can be used as a means of maximizing the survival distribution. RSF is available from the R package ”randomForestSRC”. The forest then combines the outputs of the decision tree to produce the final output. It is a bootstrapping method that randomly selects elements from the training data. This is a random process. A collective decision based on apلurality of decisions and a final decision based on majority vote.

Algorithm 1 Random survival forest method

1. **procedure** GENERATE B BOOTSTRAP SAMPLES FROM THE ORIGINAL DATA SET.
2. Construct a survival tree for each bootstrap sample in each node of the tree.
3. Choose m ≤ n, without replacement, randomly.
4. Let the tree grow to the maximum, but observe the following limits: The terminal of the increase reaches the minimum value Evaluate the observation d ≥ 0.
5. Calculate the risk factor for each tree and then average these functions.
6. Use the OOB set and calculate the prediction error.

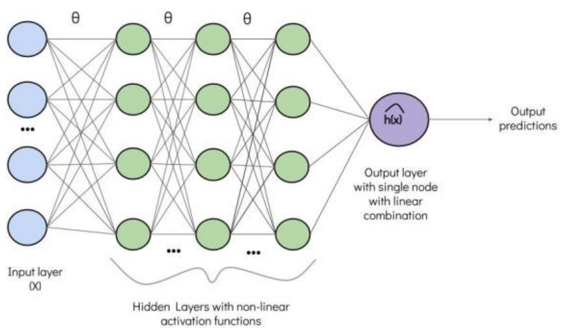

Fig. 4. DNNSurv structure

3.2. DNN Surv Model

DNN Survival Model leverages the power of deep learning, utilizing neural network architectures with multiple layers to discern intricate temporal dependencies and complex relationships within the data. This is particularly advantageous when dealing with datasets where the event of interest may be influenced by a myriad of factors in a non-linear fashion. The architecture of a DNN Survival Model typically consists of layers of interconnected nodes, each layer contributing to the model's ability to extract hierarchical features from the input data. Through the process of training, wherein the model learns from historical examples, the DNN can capture subtle patterns that might be challenging for traditional models.

DeepSurv is a deep, feed-forward neural network that allows the network to predict the risk level of each individual based on their weights. Most the structure of this neural network is very simple. Compared to the Simon-Faraghi network, DeepSurv is structured with multiple hidden layers. The input data X is represented as a set of observed covariates. The layers in this model are all connected by nonlinear activation layers, and their numbers are not necessarily the same. Each node keeps track of the destruction process. The output method has only one function line that gives the output (risk estimate). Neural network algorithms can be efficient even if they have a very simple structure (such as a single hidden layer). DNNSurv model DNN is created by combining the survival model and the Cox PH model and can be used for HD survival data. Deep Neural Networks are commonly known as DNNSurv and are used specifically for survival analysis Neural network designed to predict survival outcomes for the task of survival analysis for model delays, failures or other events on the data bus, for example setting the pattern.

Identify relevant features in the bus transportation dataset, such as departure times, distances, and other factors that may influence the survival times or events of interest. Define the target variables, typically survival time and censoring status. Survival time represents the duration until the event of interest (e.g., delay) or censoring, and censoring status indicates whether the event has occurred or if the data is censored.The input layer of the neural network takes in the selected features.One or more hidden layers are employed to capture complex patterns and relationships within the data. Activation functions like Relu (Rectified Linear Unit) are commonly used to introduce non-linearity. Where Hidden layer as 32 Neurons of 1*1 Dimensions.The output layer typically has a linear activation function to produce continuous predictions representing the estimated survival times. The Output Layer 256 Units. Where the output layer with Sigmoid Value. Loss Function and Optimization Selection holds significant importance survival analysis commonly involves considering specific survival parameters like mean squared error. Employing an optimizer such as Adam optimizer aids in minimizing the surplus selected during training. Adam optimizer functions as a variable ratio calculator, computing individual ratios for different factors. It has the capability to modify the stochastic gradient descent method and readjust the network weights based on the available information. Format the input data and output labels to feed into the DNN. Input features and survival times are used for training.

$$m_t = \beta m_t - 1 + (1-\beta)\frac{\partial l}{\partial w_t} v_t = \beta_2 v_t - 1 + (1-\beta_2) * [\frac{\partial l}{\partial w_t}]^2 \quad (2)$$

where W_t is Weights at time t. $W_t + 1$ Weights at time t+1. ∂L is a Derivative of Loss Function. ∂W_t Derivative of weights at time t . V_t Sum of the square of past gradients. (initially, $V_t = 0$) . β Moving average parameter (const,
0.9).

The model is trained over several periods during which it learns the lifelong concept map. Evaluate the trained DNNSurv model using appropriate survival metrics. The consistency index (C-index) is often used to measure the classification ability of the model, and the table below shows the set of hyperparameters used in training and optimization. Since there is no built-in hyperparameter finding option in the Pycox suite, this parameter is sent manually. DNNSurv is specially designed to control data censorship, this situation will not occur at the end of the observation period. To achieve the best results, effective methods and early termination are used to monitor quality and prevent overwork during training. DNNSurv is a unique neural network architecture designed for survival analysis in transportation system. By using its ability to learn complex patterns from data, DNNSurv becomes a powerful tool for predicting lifespan, efficiency and prediction of events such as delays or malfunctions in transportation.

Algorithm 2 DnnSurv Method

1. **procedure** LOAD YOUR BUS TRANSPORTATION DATASET.
2. Split the dataset into training and testing sets.
3. Design a neural network architecture for survival analysis Four layer to collect filtered features from DNNlayer. adding 32 nodes per layer.
4. convert multidimension data to single dimension data (Flatten ()) defining output layer add (Dense(units = 256))
5. Compile Adam optimizer and Loss Function.
6. DNNSurv is assessed using metrics by the C-index and Auc Score.

3.3. C-Index

The Concordance index, often called the C-index or C-statistic, is a measure of performance. In survival analysis or time-to-event model. Simply put, it measures the discriminatory power of a model, that is, the ability of the model to distinguish between individuals who experienced an event earlier and individuals who experienced that event later. It assesses the ability of a predictive model to correctly order pairs of observations with respect to their existence Pairwise comparisons of times For each pair of individuals in the dataset, the time of existence of the two individuals A comparison is made. A concordant pair is considered concordant if the individual has a longer observed survival time Also has a longer predicted survival time according to the model. If Individuals with longer observed survival times have shorter model-predicted survival times The c-index is then calculated as the ratio of congruent pairs to the sum of discordant pairs. C-index to evaluate the discriminative power of different models or to compare the performance of different models. Concurrence in the prediction of survival outcomes. Higher C-index values indicate better predictive performance.

$$c = p(\hat{T}1 > \hat{T}2 | T1 > T2) \tag{3}$$

where $\hat{T}1$ and $\hat{T}2$ the approximate times of existence of T1 and T2 respectively, which can be obtained by Estimates of risk or prognostic scores g(x;β)

3.4. Area Under the Curve (AUC)

The Time-dependent area under the curve (AUC) is a metric used Performance of a binary classification model, often in terms of receiver operating characteristic (ROC) curves. In survival analysis, it is used to evaluate the accuracy of the model In the distinction between individuals who experience the phenomenon of interest and those who do Not more than a certain period. AUC values range from 0 to 1, where a value of 0.5 indicates random chance and a value of 1.0 indicates perfect discrimination. A high AUC indicates that Better overall performance of the model in terms of sensitivity and specificity. The AUC curve provides a concise and intuitive summary of a model's ability to classify Examples In the context of survival analysis for bus transportation datasets, AUC curves can be performed used to assess how well the model discriminates between different survival outcomes, viz As a delay or prediction of other events. A higher AUC indicates a more effective model in terms of Discrimination provides insight into the overall performance of the model across different Threshold values.

$$Se(t) = Sensitivity(c,t) = p|M > C|T \leq T; \quad SP(t) = Sensitivity(c,t) = p|M \leq C|T < T \tag{4}$$

4. Results

The proposed system evaluates the performance of the DNNSurv and RSF models using two evaluation mea- sure. Integrating survival analysis, random forests, and DNN models on a bus traffic dataset it aims not only to provide a detailed understanding of service reliability but also presents a benchmark predictive capabilities of

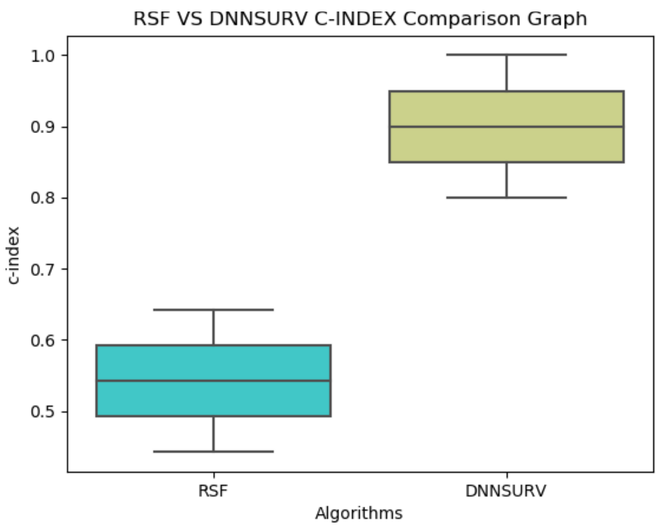

Fig. 5. Boxplots of the C-index for Two models

these advanced models. C-index and AUC serve as critical measures, offers a robust framework for evaluating the performance of each model and guidance for choosing the most an effective approach to QoS prediction in bus traffic. The C-Index measures discriminatory power such models demonstrate how well they can distinguish between individuals experiencing a given event in different locations time points. It quantifies the probability that, given two randomly selected individuals, the model predicts correctly dictates the order in which they experience the event. The index ranges from 0.5 (indicating random chance) to 1.0 (perfect discrimination). The ROC curve plots the true positivity rate (sensitivity) against the false positive rate (1-specificity) for different thresholds and AUC scores provides a single, comprehensive measure discriminative power of the model. A higher AUC score, closer to 1, indicates better model performance excellent ability to correctly classify instances. Conformity Index (Cindex), also known as the C-statistic, is a measure commonly used in survival analysis and binary classification to evaluate predictive accuracy and Model. It assesses how well the model discriminates between pairs of samples based on their predicted outcomes.

C-index and AUC as main criteria provide a strong basis for evaluating the performance of each model and

guide the selection of the best method for initial QoS estimation in the bus. Figure 5 shows the prediction performance of two models (i.e., DNNSurv, RSF) for C-indexing based on data files. The consistency index plays an important role in assessing the accuracy and reliability of model predictions and provides researchers and practitioners with a quantitative measure of the model's ability to match different outcomes of interest. The area under the curve(AUC)score is a widely used metric in machine learning and statistics especially regarding binary classification problems. It works as a benchmarking tool to evaluate the effectiveness of classification models. From Figure 5, we can see that the DNNSurv model provides the best performance compared to the RSF model. Where RSF C-index value is 0.64, C-Index value of DNNSurv model is 0.99. It can be concluded that, especially for the dataset, the DNNSurv model performs better than other models (e.g. RSF) in terms of C-index. The fact that both parameters are close to 1.0 means that the model has a very good solution and is effective. If a measurement is higher than the previous one, it will indicate that the model is performing well. Figure 6 shows the prediction performance of two models(e.g. DNNSurv, RSF) based on the dataset-based AUC curve. As can be seen in Figure 6, the DNNSurv model provides the best performance compared to the RSF model. AUC value RSF is 0.1, Auc value of DNN model is
1.4. We can also conclude that for the dataset the DNNSurv model performs better than other models (e.g. RSF) in terms of the AUC curve. As a result, it can be seen from Figure 5 and Figure 6 that the overall performance of the DNNSurv model is better than the RSF model in terms of both dataset metrics. When the given delay is estimated, the values are 1.62 and 157 ms, respectively. Due to the censored nature of survival data, calculating AUC in survival analysis is more difficult than the usual binary distribution problem. Specialized statistical methods and software packages such as the Survival ROC package often provide an opportunity for the AUC calculation to work. Time-dependent AUC can be used to evaluate and compare the predictive accuracy of different survival models or to quantify the additive value of the variable in predicting survival. Result over time.
Calculating AUC in survival analysis is more complex than in traditional binary classification problems dueto the censored nature of survival data. Specialized statistical methods and software packages such as R with

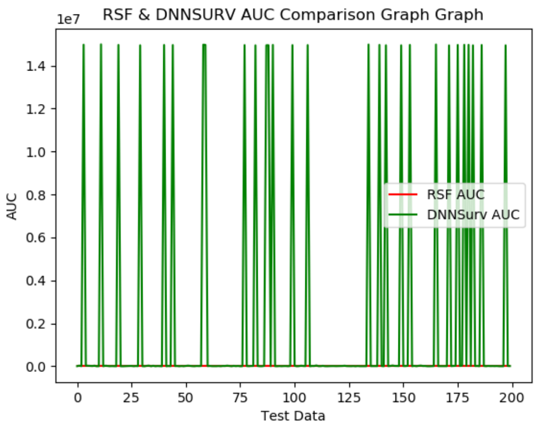

Fig. 6. The time-dependent AUC for Two models

packages such as survival-ROC or time-ROC often provide functions to calculate time-dependent AUC. The time-dependent AUC is useful for evaluating and comparing the predictive accuracy of different survival models or evaluating the added value of specific variables in predicting survival outcomes over time.

5. CONCLUSION

The utilization of Survival Analysis employing Random Survival Forest (RSF) and Deep Neural Network (DNN) survival models on a dataset related to bus transportation, utilizing the 5-fold cross-validation (5-CV) method for C-index and time-dependent scores, yields valuable insights into the predictive performance of these models. The RSF model exhibits a favorable average C-index, indicating its ability to effectively predict survival times. Furthermore, the average time-dependent AUC score highlights the model's proficiency in ranking survival times at different intervals. On the other hand, the DNN survival model demonstrates a competitive average C-index, suggesting its effectiveness in predicting survival outcomes. The time-dependent AUC scores further validate the model's adaptability and strong performance across various stages of the survival curve.

The selection between RSF and DNN depends on the specific characteristics of the dataset and the modeling requirements. RSF offers interpretability and robustness, while DNN excels in capturing intricate relationships within large datasets. The utilization of 5-fold crossvalidation ensures a reliable evaluation of model generaliza- tion, mitigating potential overfitting concerns and involving hyperparameter tuning. By combining the RSF and DNN survival models and assessing them through 5-fold cross-validation with C-index and time-dependent AUC scores, a comprehensive understanding of their respective strengths in predicting survival outcomes within the bus transportation context is achieved.

These insights enable informed decision-making for transportation planning and management, potentially lead- ing to improved efficiency and reliability in bus services. Continuous monitoring and evaluation with new data may be necessary for future enhancements, enhancing the adaptability of the models to changing transportation patterns.

References

1. Clark, T., Bradburn, M., Love, S. et al. Survival Analysis Part I: Basic concepts and first analyses. Br J Cancer 89, 232–238 (2003). https://doi.org/10.1038/sj.bjc.6601118.
2. Bojan Kostic, Mathilde Pryds Loft, Filipe Rodrigues, Stanislav S. Borysov, ”Deep survival modelling for shared mo- bility”,Transportation Research Part C: Emerging Technologies,Volume 128,2021,103213, ISSN 0968-090X.
3. Amin Vahedian, Xun Zhou, Ling Tong, W. Nick Street, Predicting Urban Dispersal Events: A Two-Stage Framework through Deep Survival Analysis on Mobility Data ,2019,1905.01281.
4. Lili Zhao and Dai Feng,DNNSurv: Deep Neural Networks for Survival Analysis Using Pseudo Values,2020,1908.02337.
5. Stephane Fotso,Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework,2018.
6. Wang H, Li G. Extreme learning machine Cox model for high-dimensional survival analysis. Stat Med. 2019 May 30;38(12):2139-2156. doi: 10.1002/sim.8090. Epub 2019 Jan 10. PMID: 30632193; PMCID: PMC6498851.
7. S. Kamalapurkar, Savitha and G H, Samyama Gunjal, ”Online Portal for Prediction of Heart Disease using Machine Learning Ensemble Method(PrHD-ML),” 2020 IEEE Bangalore Humanitarian Technology Conference (B-HTC), Vi- jiyapur, India, 2020, pp. 1-6, doi: 10.1109/B-

HTC50970.2020.9297918.

8. Lee S, Lim H. Review of statistical methods for survival analysis using genomic data. Genomics Inform. 2019 Dec;17(4):e41. doi: 10.5808/GI.2019.17.4.e41. Epub 2019 Dec 20. PMID: 31896241; PMCID: PMC6944043.
9. Kantidakis G, Biganzoli E, Putter H, Fiocco M. A Simulation Study to Compare the Predictive Performance of Survival Neural Networks with Cox Models for Clinical Trial Data. Comput Math Methods Med. 2021 Nov 28;2021:2160322. doi: 10.1155/2021/2160322. PMID: 34880930; PMCID: PMC8646180.
10. Hao, L.; Kim, J.; Kwon, S.; Ha, I.D. Deep Learning-Based Survival Analysis for High Dimensional Survival Data. Mathematics 2021, 9, 1244. https://doi.org/10.3390/math9111244.
11. Madhuri, V. M., Gunjal GH Samyama, and Savitha Kamalapurkar. "Air pollution prediction using machine learning supervised learning approach." Int J Sci Technol Res 9.4 (2020): 118-123.
12. Rai, Sushmita , Mishra, Prabhakar, Ghoshal, Uday. (2022). Survival analysis: A primer for the clinician scientists. Indian Journal of Gastroenterology. 40. 10.1007/s12664- 021-01232-1.
13. Barakat, Ahmed, Mittal, Aaina, David, Ricketts, Rogers, Benedict. (2019). Understanding survival analy- sis: actuarial life tables and the Kaplan–Meier plot. British Journal of Hospital Medicine. 80. 642-646. 10.12968/hmed.2019.80.11.642.
14. Suresh Chavhan, Deepak Gupta, Chandana B N, Ashish Khanna, A Ram Mohan, Joel Rodrigues,"An efficient context- aware vehicle incidents route service management for intelligent transport system", IEEE Systems Journal, 2021
15. Suresh Chavhan, Pallapa Venkataram, "Prediction based traffic management in a metropolitan area", Journal of traffic and transportation engineering, 2020
16. Suresh Chavhan, Pallapa Venkataram, "Emergent intelligence based QoS routing in MANET", Procedia Computer Science, 2015