GEORGETOWN UNIVERSITY
McCourt School of Public Policy

MASSIVE DATA INSTITUTE

COXNET: A DEEP LEARNING FRAMEWORK FOR PREDICTING RARE EVENTS IN SOCIAL SCIENCES

This project explores the application of deep learning to Event History Analysis (EHA) for predicting conflict events. CoxNet, a novel model incorporating temporal dependencies and frailty, demonstrates improvements over traditional methods in handling complex social science data.

AUTHORS

Billy McGloin, M.S. Data Science & Analytics, 2025
Supervisor: Dr. Ioannis Ziogas,
Georgetown McCourt School of Public Policy

INTRODUCTION

Machine learning (ML) models have gained traction in various scientific disciplines, but remain underutilized in social sciences, particularly in political science. Traditional statistical methods such as parametric regression models dominate the field but often struggle with the challenges posed by high-dimensional, censored, and heterogeneous social science data. These limitations hinder the ability to model and predict complex social phenomena accurately.

Building on prior work, this project improves CoxNet, a deep-learning framework initially developed to address these challenges by incorporating temporal dependencies and unobserved heterogeneity. Through enhancements such as input reconstruction to handle missing data, expanded hyperparameter optimization, and the inclusion of new data sources, CoxNet further extends the capabilities of Event History Analysis (EHA). Compared to traditional methods, it offers a more robust and flexible approach for predicting conflict events, such as wars.

OUR DATA

We used data from the Correlates of War project, which includes variables like power distribution, alliances, democracy scores, and hostility levels (cwhostd) for dyads over time. To prepare the data for deep learning, we standardized variables and addressed missing data using input reconstruction with a binary masking mechanism and subject embeddings, leveraging sequential information for imputation. Synthetic datasets with random noise and predictable patterns were also created to evaluate model performance under controlled conditions. These steps ensured the data's readiness for training and a thorough evaluation of CoxNet's capabilities.

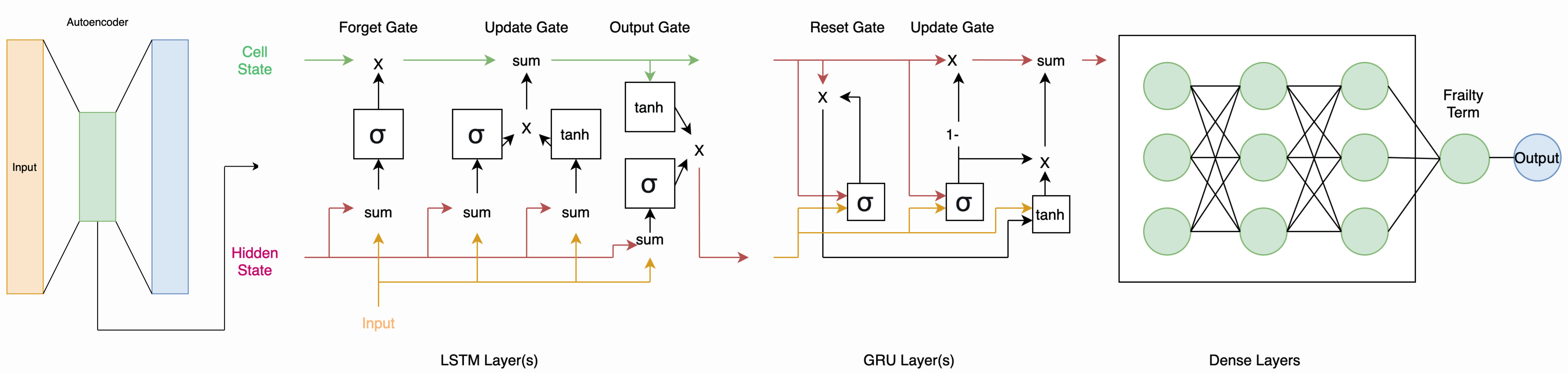
ETHICAL CONSIDERATIONS

This project uses sensitive data concerning international conflicts, raising concerns about potential misuse, such as oversimplifying complex sociopolitical events. To address this issue, we position CoxNet as a research tool, emphasize transparency in our methodology, and clearly communicate the model's limitations, particularly the challenges of handling high-dimensional, heterogeneous data with inherent biases.

MODEL DEVELOPMENT AND IMPLEMENTATION

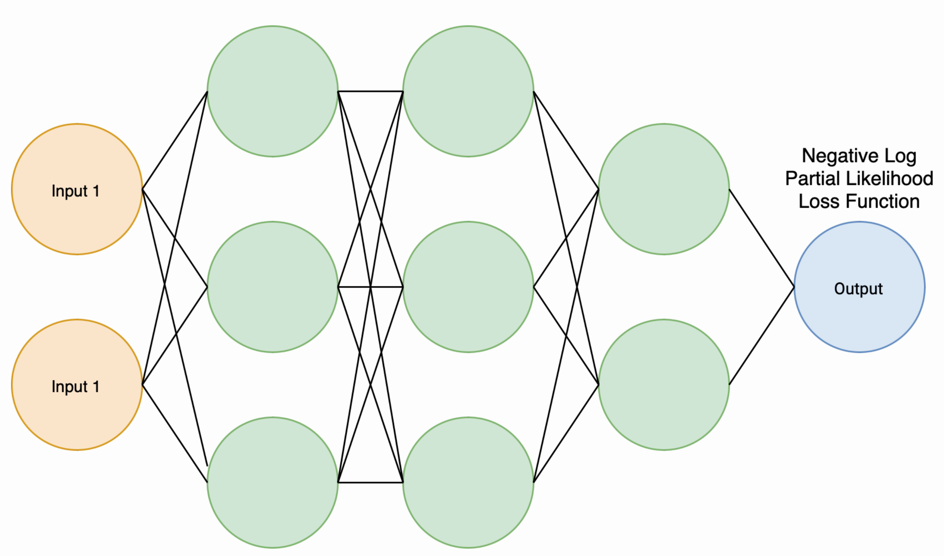
We developed and evaluated three models: DeepSurv, SurvNet, and CoxNet. DeepSurv estimates risk functions using covariates, while SurvNet incorporates multi-task learning with input reconstruction, survival classification, and Cox regression tasks. Building on these, CoxNet combines stacked RNN layers (LSTMs and GRUs) with dense layers and introduces a frailty node to capture unobserved heterogeneity. It also incorporates input reconstruction using subject embeddings and binary masking to handle missing data, ensuring compatibility with sequential and non-sequential inputs. These innovations make CoxNet a robust and flexible framework for analyzing complex social science datasets.

MODEL ARCHITECTURES



COXNET

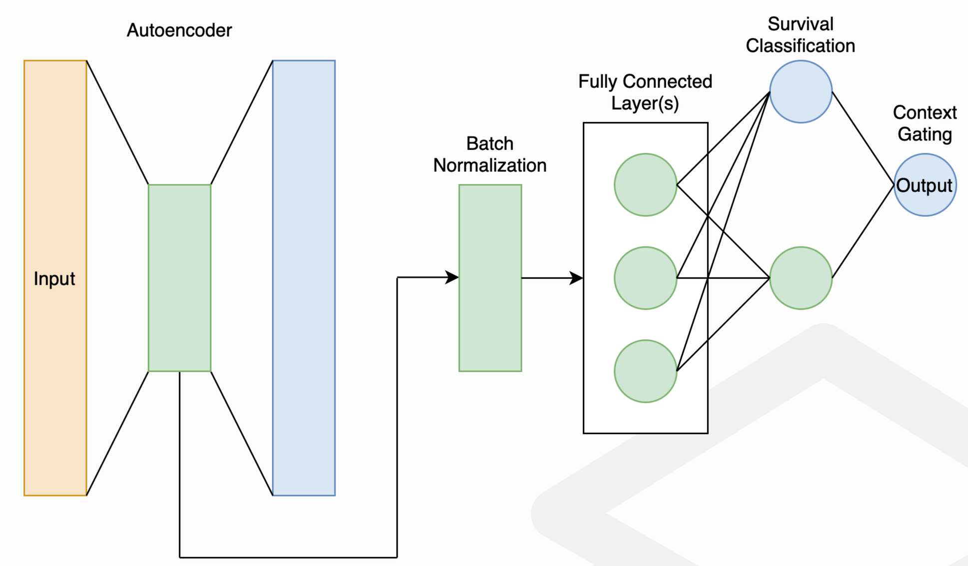
An advanced model using stacked RNN layers and a frailty node to capture temporal dependencies and unobserved heterogeneity in event prediction.



DEEPSURV

A feedforward neural network that estimates risk functions using covariates, employing a negative log partial likelihood loss function.

Source: Katzman, J.L., Shaham, U., Cloninger, A. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 18, 24 (2018).



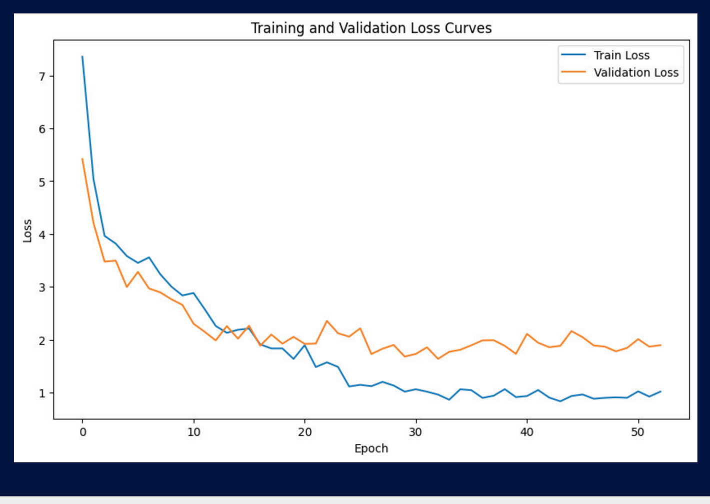
SURVNET

A multi-task learning framework that handles incomplete data through input reconstruction, survival classification, and Cox regression.

Source: Wang J, Chen N, Guo J, Xu X, Liu L, Yi Z. SurvNet: A Novel Deep Neural Network for Lung Cancer Survival Analysis With Missing Values. Front Oncol. 2021 Jan 20;10:588990. doi: 10.3389/fonc.2020.588990. PMID: 33552965; PMCID: PMC7855857.

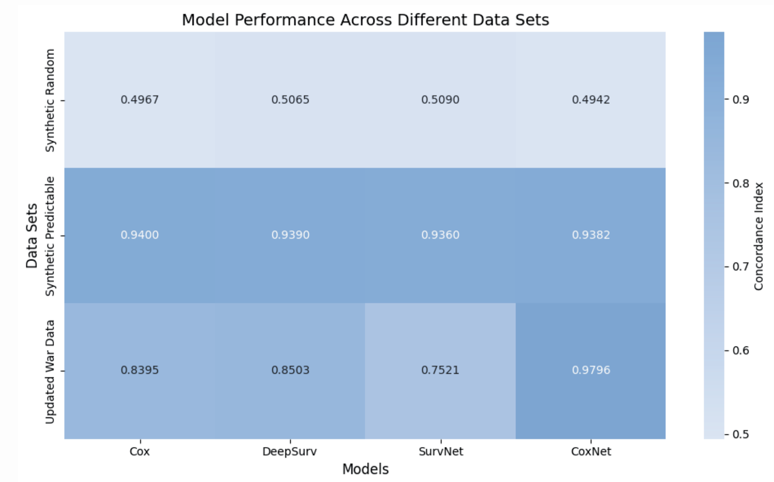
MODEL TRAINING

CoxNet was trained using a random hyperparameter search to optimize performance across the expanded dataset. The training process incorporated input reconstruction and enhanced architectural components, ensuring the model could effectively handle missing data and capture complex temporal dependencies. The visualization below shows the training and validation loss curves for CoxNet using the updated data, highlighting its improved learning dynamics and stability.



MODEL PERFORMANCE

CoxNet demonstrated competitive performance on synthetic datasets, achieving results nearly identical to other models, while significantly outperforming both traditional Cox models and other deep learning frameworks on the updated war data. The table below highlights each model's concordance index (c-index) across synthetic random data, synthetic predictable data, and updated war data. CoxNet's performance on the war data underscores its exceptional ability to handle complex, real-world social science datasets.



Comparative C-index Values:

Baseline Cox - 0.8395
SurvNet - 0.7521
DeepSurv -0.8503
CoxNet - 0.9796

CONCLUSION AND FUTURE DIRECTIONS

CoxNet has demonstrated it's potential as a powerful framework for predicting rare events in social sciences, particularly in conflict studies. By incorporating temporal dependencies, modeling unobserved heterogeneity through a frailty node, and addressing missing data with input reconstruction, CoxNet significantly outperformed traditional Cox models and other deep learning frameworks on real-world datasets. Its versatility in handling sequential and non-sequential data highlights its value as a flexible and robust tool for analyzing complex social phenomena. Looking ahead, our focus will be on several key areas of improvement and expansion:

- **Broader Applications:** Apply CoxNet to new domains within social sciences to assess its generalizability and uncover novel insights.
- **Enhancing the Architecture:** Explore alternative RNN designs, such as custom LSTM layers, or integrate transformer-based components to further improve CoxNet's ability to capture temporal patterns.

This project lays the groundwork for future research that can leverage advanced ML techniques to address complex and nuanced questions in social sciences. By continuing to refine our model and broaden its application, we hope to contribute valuable tools and methodologies for researchers and policymakers alike.