Enhancing Prognosis of Chronic Kidney Disease -Related Outcomes through Multimodal Deep Learning Approaches

Yawen Wei, Nilay Bhatt, Kiara Zhang, Shangyun Zhangliang

ABSTRACT

Chronic kidney disease (CKD) presents a significant global health burden, leading to substantial annual mortality rates. Timely detection and precise prognosis are pivotal for effective CKD management. In this project, we employ a multifaceted deep learning approach to enhance the accuracy of CKD prognosis. Our investigation encompasses a spectrum of models and methodologies, ranging from logistic regression, XGBoost, to LSTM, BERT, and BART models. Leveraging the MIMIC-III dataset, comprising de-identified health records from intensive care units, we amalgamate relevant tables to construct a comprehensive dataset. We adhere to the OMOP CDM standards for data standardization, encompassing structured values alongside unstructured variables.

Noteworthy insights include the balanced performance observed with ElasticNet and BERT, as well as the precision-oriented nature of XGBoost and BART models. While LSTM exhibits promise, its efficacy across certain CKD classes may necessitate further optimization. Evaluation of our models encompasses diverse metrics such as training loss, validation accuracy, and AUC curves, providing a comprehensive assessment of performance.

Keywords: *Chronic Kidney Disease (CKD), Deep Learning, MIMIC-III Dataset, mortality, transformers*

INTRODUCTION

a. Background

Chronic kidney disease (CKD) is a progressive disorder that gradually diminishes the functioning of the kidneys, often without any noticeable symptoms until later stages. As CKD advances, it can cause symptoms such as fatigue, vomiting, and cognitive disorientation. Complications include high blood pressure, bone disorders, and heightened cardiovascular risks.

CKD arises from various causes such as diabetes, hypertension, glomerulonephritis, and polycystic kidney disease, with a familial predisposition.

Due to its substantial impact on global health, resulting in 5 to 10 million deaths annually due to kidney-related complications, addressing chronic kidney disease (CKD) is paramount. Treatment options for CKD revolve upon diagnosing the underlying cause and may include lifestyle modifications, medication, dialysis, or kidney transplantation. Early detection and intervention are crucial in halting the advancement of kidney disease and enhancing patient outcomes. Nevertheless, precisely forecasting the evolution of chronic kidney disease (CKD) and implementing successful interventions present difficulties, as numerous instances remain unnoticed until advanced stages inside healthcare systems.

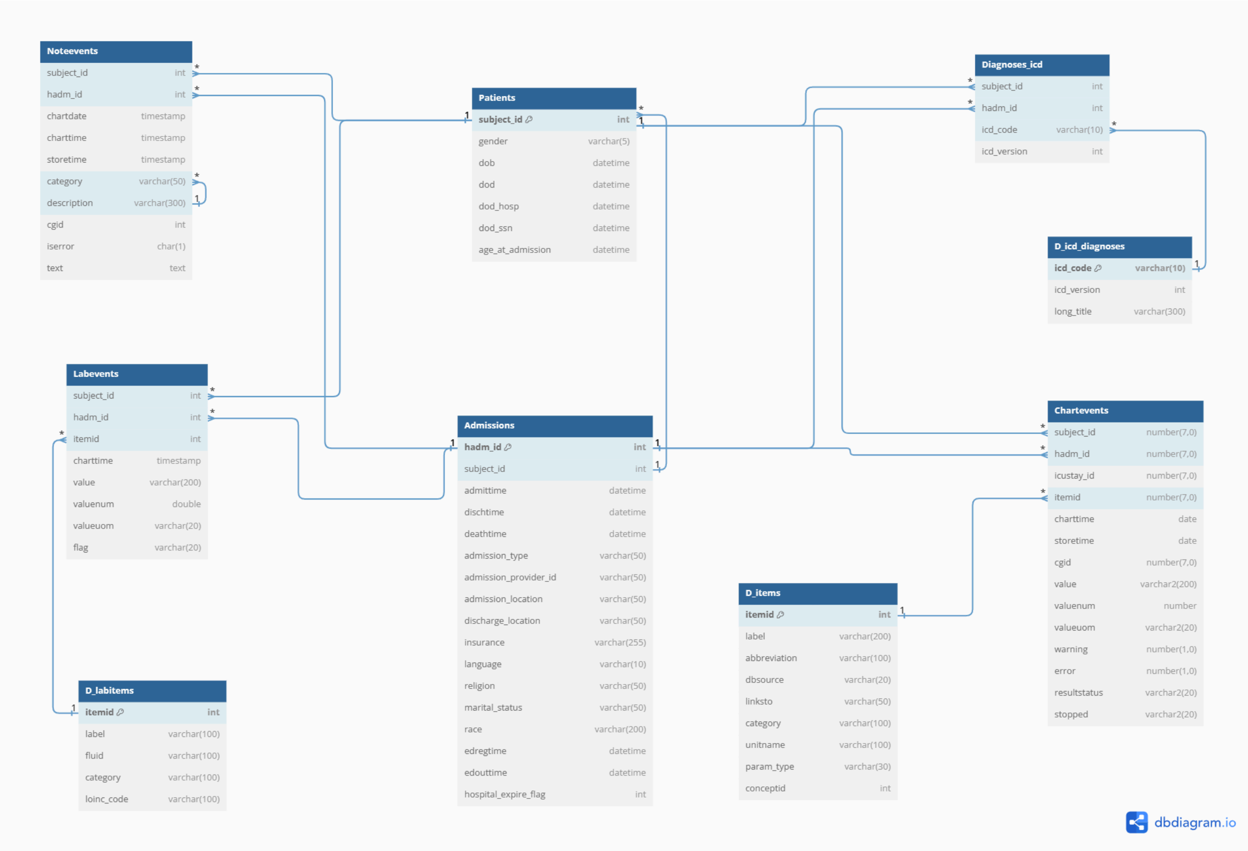
Presently, CKD models utilize several machine learning techniques such as logistic regression, random forest, XGBoost, deep neural networks, K-Nearest Neighbor, and support vector machines. Every model provides its advantage for identifying CKD at an early stage, predicting its course, and tailoring treatment plans to individual patients.

To address these challenges,our project seeks to employ multimodal deep learning methods to improve the prediction of outcomes associated with CKD. Through the analysis of data obtained from the MIMIC-III database, our objective is to find models that exhibit the greatest possible degree of accuracy to be able to enhance the prognosis of CKD and tailor patient management accordingly.

b. Data Preparation

For this study, we utilize the Medical Information Mart for Intensive Care III (MIMIC-III) database, which is a publicly available resource with comprehensive data from intensive care units at Boston's Beth Israel Deaconess Medical Center from 2001 to 2012. This dataset includes 1,159 top-level ICD-9 codes and 112,000 clinical reports pertaining to 46,520 patients.

To conduct our analysis, we utilized nine crucial tables from the MIMIC-III database: Patients, Admissions, Diagnosis_icd, D_icd_diagnosis, Chartevents, D_items, Labevents, D_labitems, and Noteevents. We extracted relevant data from the MIMIC-III database using Google BigQuery, joining multiple tables using common identifiers such as subject_id, hadm_id, item_id, and icd9_code. The Patient table offers individual patient details, distinguished by the unique subject_id, while the Admission table logs each hospitalization with the unique hadm_id identifier. The Diagnosis_icd table contains diagnoses, using ICD-9 codes, for patients (subject_id) based on their unique admission to the hospital (hadm_id). Labevents detail laboratory-based measurements, while Chartevents display vital signs and additional nursing details. Lastly, Noteevents provides comprehensive clinical notes. All these tables are connected to specific patients and hospital admissions via subject_id and hadm_id. Additionally, we utilize SQL techniques to filter and extract patients diagnosed with chronic kidney disease based on ICD-9 codes.



Noteevents	
subject_id	int
hadm_id	int
chartdate	timestamp
charttime	timestamp
storetime	timestamp
category	varchar(50)
description	varchar(300)
cgid	int
iserror	char(1)
text	text

Patients	
subject_id	int
gender	varchar(5)
dob	datetime
dod	datetime
dod_hosp	datetime
dod_ssn	datetime
age_at_admission	datetime

Diagnoses_icd	
subject_id	int
hadm_id	int
icd_code	varchar(10)
icd_version	int

D_icd_diagnoses	
icd_code	varchar(10)
icd_version	int
long_title	varchar(300)

Labevents	
subject_id	int
hadm_id	int
itemid	int
charttime	timestamp
value	varchar(200)
valuenum	double
valueuom	varchar(20)
flag	varchar(20)

Admissions	
hadm_id	int
subject_id	int
admittime	datetime
dischtime	datetime
deathtime	datetime
admission_type	varchar(50)
admission_provider_id	varchar(50)
admission_location	varchar(50)
discharge_location	varchar(50)
insurance	varchar(255)
language	varchar(10)
religion	varchar(50)
marital_status	varchar(50)
race	varchar(200)
edregtime	datetime
edouttime	datetime
hospital_expire_flag	int

Chartevents	
subject_id	number(7,0)
hadm_id	number(7,0)
icustay_id	number(7,0)
itemid	number(7,0)
charttime	date
storetime	date
cgid	number(7,0)
value	varchar2(200)
valuenum	number
valueuom	varchar2(20)
warning	number(1,0)
error	number(1,0)
resultstatus	varchar2(20)
stopped	varchar2(20)

D_items	
itemid	int
label	varchar(200)
abbreviation	varchar(100)
dbsource	varchar(20)
linksto	varchar(50)
category	varchar(100)
unitname	varchar(100)
param_type	varchar(30)
conceptid	int

D_labitems	
itemid	int
label	varchar(100)
fluid	varchar(100)
category	varchar(100)
loinc_code	varchar(100)

Figure. Database Schema: ER Model

The dataset was standardized using the Observational Medical Outcomes Partnership (OMOP) schema, resulting in structured and unstructured data components for comprehensive analysis. The structured data includes numeric values from labs, vitals, and patient demographics, along with the critical target variable "hospital_expire_flag" indicating in-hospital mortality. Conversely, the unstructured component captures clinical text alongside associated labels, providing valuable insights into patient care and outcomes.

In the structured data, each row represents one hospital encounter, totaling 232 rows and 56 variables after dropping columns with excessive missing values (≥20%). Conversely, each row in unstructured data represents one mission which corresponds to every clinical note made in that mission. In this case, the unstructured data contains 25700 entries. The preprocessing of text began by converting all text to lowercase to standardize the data, followed by removing all punctuation and non-alphanumeric characters except spaces using a regular expression. The cleaned text was then tokenized into individual words, and common stopwords were removed to focus the analysis on more meaningful words. Subsequently, tokens that were purely numeric were discarded, and the remaining tokens were lemmatized to their base or dictionary forms to consolidate information. These steps transformed the entire column of raw text into a column of cleaned and tokenized text. Here we have 3 columns in unstructured data, hadm_id, hospital_expire_flag, and preprocess_text.

METHODS

a. EDA

Exploratory Data Analysis (EDA) was conducted to provide a foundational understanding of the dataset's characteristics and the relationships between various clinical parameters. The primary goal of EDA is to uncover underlying patterns, spot anomalies, test assumptions, and check the validity of the data, ensuring informed model development and hypothesis testing.

Descriptive statistics were computed for both continuous and categorical variables to summarize central tendencies, variability, and distributions. For continuous variables, measures such as mean, median, standard deviation, minimum, and maximum values were calculated. For categorical variables, frequencies and proportions were used to understand demographic distributions such as gender and ethnicity. These statistics are crucial for assessing the data’s structure and guiding further data preprocessing steps. Histograms were used to visualize the distribution of each variable, helping to identify skewness, outliers, and the general shape of the data distribution. Boxplots supplemented this analysis by providing a visual representation of the distribution's quartiles and outliers. These visualizations are critical for detecting anomalies and understanding the spread and central tendency of the data.

The rationale for employing these specific EDA techniques stems from their effectiveness in revealing the data's nature and quality without making any assumptions about its distribution or underlying relationships. This approach is particularly pertinent in medical datasets where variable interactions can be complex and predictions can significantly impact patient outcomes.

b. Feature Selection

Mutual information is a statistical measure that quantifies the amount of information one variable contains about another. This metric is particularly useful in feature selection for predictive modeling because it evaluates the reduction in uncertainty for one variable given the knowledge of another, which in this context is the outcome variable, hospital mortality. We employed mutual information to identify and prioritize features from the MIMIC-III dataset that have the strongest associations with hospital mortality. This approach ensures that the predictive models we develop are focused on the most relevant variables, thereby enhancing the models' effectiveness in forecasting outcomes based on the identified key predictors.

c. Model Selection

Logistic regression
We decided to create a logistic regression model as a baseline model. To prepare the dataset for logistic regression analysis, categorical variables, including ‘gender’ and ‘ethnicity’, were first transformed into a format suitable for modeling. This was achieved by converting categorical variables into dummy variables. For this purpose, the Python library pandas was utilized, employing the get_dummies function. This function transforms each categorical feature into multiple binary variables, each representing a category in the original feature. The parameter ‘drop_first=True’ was used to avoid multicollinearity, by dropping the first category which becomes the reference category. Furthermore, since the logistic regression model does not accept missing values, the ‘NaN’ values in numerical columns were replaced with the mean of their respective columns. This imputation method was chosen for its simplicity and effectiveness in maintaining the distribution of the variable.

The preprocessed data was used to fit a logistic regression model, which was implemented using the ‘LogisticRegression’ class from the ‘sklearn.linear_model’ module in Python. The model was instantiated with default parameters and fitted to the training data. This step involved the optimization of the model parameters to best predict the binary target variable.

ElasticNet Regularization
Regularization reduces reliance on any specific independent variable by incorporating a penalty term into the loss function. This addition helps prevent the coefficients of the independent variables from reaching extreme values. Prior to modeling, the dataset underwent a feature scaling process to standardize the variables, ensuring that each feature contributes equally to the model's prediction performance. This step is crucial, since the logistic regression models are sensitive to the scale of input features. The ‘StandardScaler’ from the ‘sklearn.preprocessing’ module was applied for scaling the data such that each feature has a mean of zero and a standard deviation of one.

To optimize our logistic regression model, a grid search was conducted to fine-tune the hyperparameters, specifically focusing on the hyperparameter ‘C’, which refers to regularization strength, and ‘l1_ratio’, which refers to the balance between L1 and L2 regularization. The grid search aimed to explore various combinations of these parameters to determine the most effective setting for the ElasticNet regularization. The ‘GridSearchCV’ tool from ‘sklearn.model_selection’ was employed, which automates the process of fitting models across a range of parameter values and selecting the combination that performs best based on cross-validation results, which ‘C’ equals to 0.1 and ‘l1_ratio’ equals to 0.25. Following the grid search, the best-performing model was selected and then used to predict the outcomes on the test dataset.

XGBoost

XGBoost was selected for its performance efficiency, scalability, and its ability to handle sparse data with missing values. Its gradient boosting framework constructs an ensemble of decision trees in a sequential manner, where each new tree corrects errors made by the previous ones. XGBoost is particularly effective for our needs due to its method of dealing with missing values; it can inherently learn the best direction to classify these values at each decision tree split during training, eliminating the need for preliminary imputation, which can introduce bias. Additionally, XGBoost includes L1 and L2 regularization to help prevent overfitting, a common challenge in high-dimensional datasets like ours. The algorithm also stands out for its computational efficiency, a critical feature when handling large datasets with numerous features. This method is well-suited for our binary classification task of predicting mortality (hospital_expire_flag).

LSTM

The Long Short-Term Memory (LSTM) model was employed to analyze unstructured clinical text data, capitalizing on its ability to capture long-term dependencies in sequence data, which is critical for understanding the context and nuances in medical notes. In parallel, a structured neural network model was designed to process numerical and categorical clinical data, allowing for a comprehensive analysis of multimodal data sources. This multimodal approach aims to integrate diverse data types to enhance predictive accuracy in assessing patient mortality.

The structured data underwent several preprocessing steps to ensure it was suitable for modeling. Missing values in numeric data were imputed using the mean, which helps in maintaining statistical integrity without introducing bias. Numeric features were scaled using standardization to ensure that the model was not biased towards variables with larger scales. Categorical variables were transformed using one-hot encoding to convert them into a machine-readable format, allowing the model to effectively interpret and learn from these data. For the unstructured text data, the following preprocessing steps were undertaken: Text data was tokenized by converting text into sequences of integers, where each integer represents a unique word in the data. This step is crucial for preparing text for input into the LSTM model. Sequences were padded to ensure that all input data had the same length, necessary for batch processing in neural networks.

Model Architecture:

- Structured Neural Network: A simple feedforward neural network with ReLU activation was used to process the structured data. This component transforms the structured input into an embedded output that can be combined with the LSTM output.
- LSTM Network: The LSTM network processes the sequential text data, capturing the important features from the input sequences with the capability to remember or forget information over long sequences.
- Integration Layer: Outputs from the structured neural network and LSTM were concatenated and passed through a fully connected layer to perform the final prediction. This layer combines the learned representations from both structured and unstructured data to make a comprehensive prediction.

The training involved several epochs where each batch of data was processed through both the structured and text networks. The model parameters were optimized using the Adam optimizer, a popular choice for deep learning tasks due to its efficient computation and adaptive learning rate features. Binary cross-entropy loss was used as the loss function, suitable for the binary classification task of predicting patient mortality. The model's performance was evaluated using several metrics, including accuracy, precision, recall, F1-score, and the area under the ROC curve (AUC). These metrics provide a comprehensive understanding of the model's effectiveness across various aspects of classification performance.

The models were trained and tested on a computing setup equipped with GPUs to expedite the computation, leveraging PyTorch as the deep learning framework for its dynamic graph construction and efficient memory usage.

BERT

The 'emilyalsentzer/Bio_ClinicalBERT' model was selected due to its pre-training on biomedical text, making it particularly suited for clinical applications. The BertTokenizer was employed to process the text data. This tokenizer converts text into a format amenable to BERT models by:

- Adding special tokens to signify the start and end of each text.
- Padding or truncating each text sequence to a consistent length of 64 tokens.
- Generating attention masks to help the model focus on relevant parts of the sequence.

The processed datasets were split into training and validation sets, with 90% of the data allocated for training and the remaining 10% used for validation. The ‘BertForSequenceClassification’ model was initialized with two labels corresponding to the binary outcome of patient mortality. Training was conducted over 5 epochs, where each epoch consisted of:

Training Loop: Each batch of data was loaded onto the GPU for faster processing, followed by a forward pass, loss computation, and a backward pass to update the model's weights. Gradient clipping was implemented to prevent exploding gradients, ensuring stable training dynamics.

***Validation Loop**:* Post-training, the model's performance was evaluated on the validation set. Accuracy was calculated to assess model effectiveness.

The training process utilized the AdamW optimizer, with a learning rate of 2e-5 and an epsilon value of 1e-8 to stabilize training. A linear scheduler with warmup was used to adjust the learning rate dynamically based on training progress. Model performance was monitored by calculating the accuracy of predictions against the actual labels during the validation phase.

BART

To address the complexity of integrating structured clinical data and unstructured text data for the prediction of hospital mortality, we adopted the Bidirectional and Auto-Regressive Transformer (BART) model. This approach leverages the advanced capabilities of BART in processing natural language while also accommodating structured data inputs.

The dataset comprised both structured and unstructured data, necessitating preprocessing steps tailored to each data type. Numeric and categorical features underwent imputation (mean for numeric and a constant value for categorical missing data) and normalization (using StandardScaler for numeric features) or encoding (OneHotEncoder for categorical features). Text data was processed using the BART tokenizer, which handles tokenization, sequence padding, and attention mask generation, preparing the text for efficient processing by the BART model. The structured and unstructured data were combined into a single dataframe, aligning by the hospital admission ID (hadm_id). This integrated dataset facilitated the simultaneous processing of textual and numeric inputs, critical for leveraging multimodal data insights.

Model Architecture:

- BART Encoder: Processes the unstructured text inputs to capture the semantic relationships within the clinical notes.
- Feature Concatenation: Outputs from the BART encoder are concatenated with the processed structured data, creating a comprehensive feature set that includes insights from both data modalities.
- Classification Layer: A fully connected layer with dropout regularization follows the concatenated features to produce the final binary classification output.

The training of the hybrid BART model was conducted over several epochs, utilizing the AdamW optimizer with a learning rate of 5e-5. This setup is chosen for its effectiveness in handling sparse gradients and its adaptability to different parameter scales, which is beneficial in multimodal learning environments. Cross-entropy loss was utilized to calculate the error between the model's predictions and the actual outcomes. This loss function is particularly suited for binary classification problems, as it quantifies the difference between two probability distributions.

Model performance was evaluated using standard classification metrics including accuracy, precision, recall, F1-score, area under the ROC curve (AUC), and Matthews correlation coefficient (MCC). These metrics provide a comprehensive assessment of the model's predictive accuracy and its ability to handle class imbalances.

The model was trained and evaluated on a computing setup equipped with GPUs to facilitate the intensive computations required by deep learning models, particularly those involving transformers like BART.

RESULTS

a. EDA

The exploratory data analysis provided crucial insights into the patient population's characteristics and the relationships between various clinical parameters. These findings are instrumental for subsequent predictive modeling and hypothesis testing within the scope of the study on hospital mortality. Understanding these distributions and correlations helps in identifying potential biases or confounding factors that may influence the outcomes of further analyses.

The exploratory data analysis of the MIMIC-III dataset highlighted significant variability in several clinical measures across the patient cohort. Age at admission showed a broad distribution with a mean of 64.05 years, reflecting a predominantly elderly patient population. The duration of hospital stay varied significantly, with a mean of 13.84 days, indicating varied severity and treatment lengths among the patients. Laboratory measures showed diverse distributions. For example, hemoglobin levels, critical for assessing anemia, had a mean value of 9.91 g/dl, slightly below normal ranges, suggesting a prevalence of anemia among the cohort. The distribution of white blood cell count, a marker of infection or inflammation, was right-skewed with a mean of 10.42 K/uL, pointing towards acute inflammatory or infectious processes in many patients.

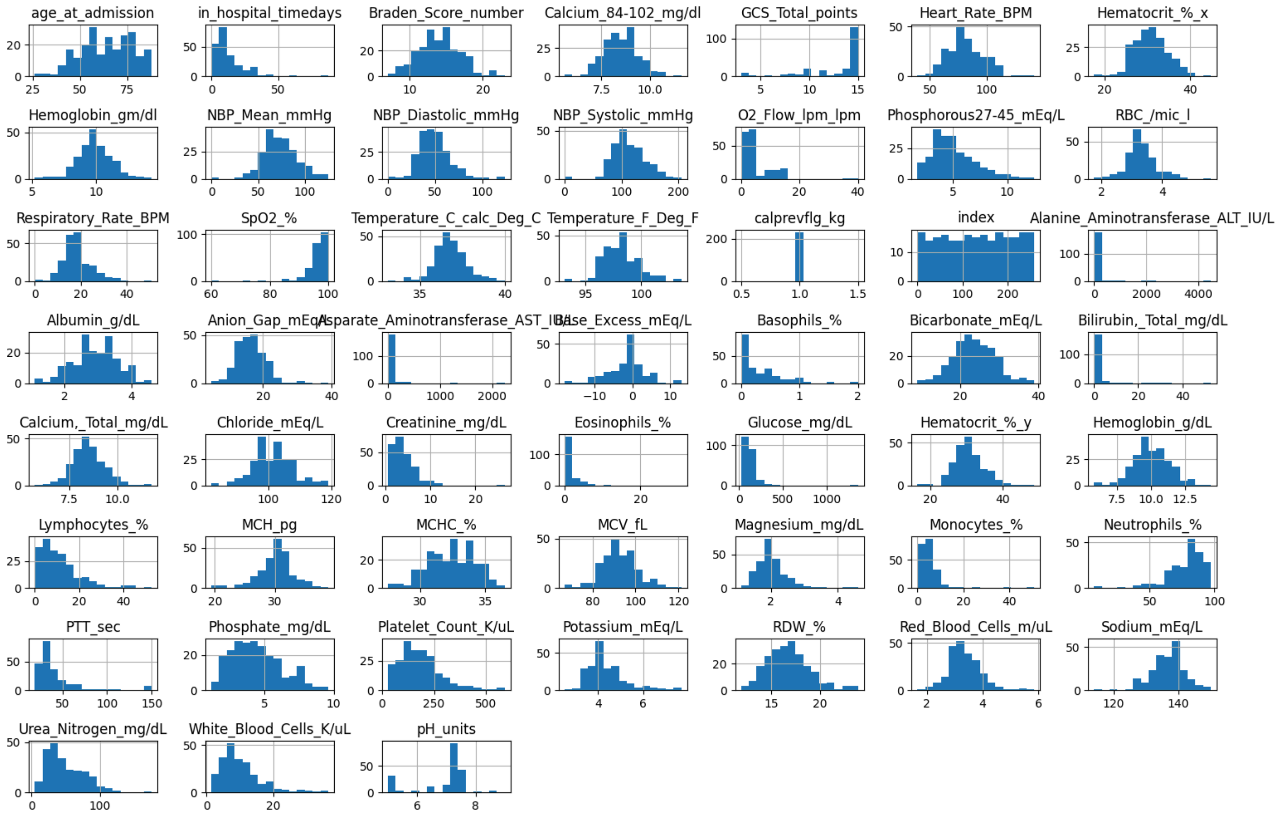


Figure: Histograms of feature distributions

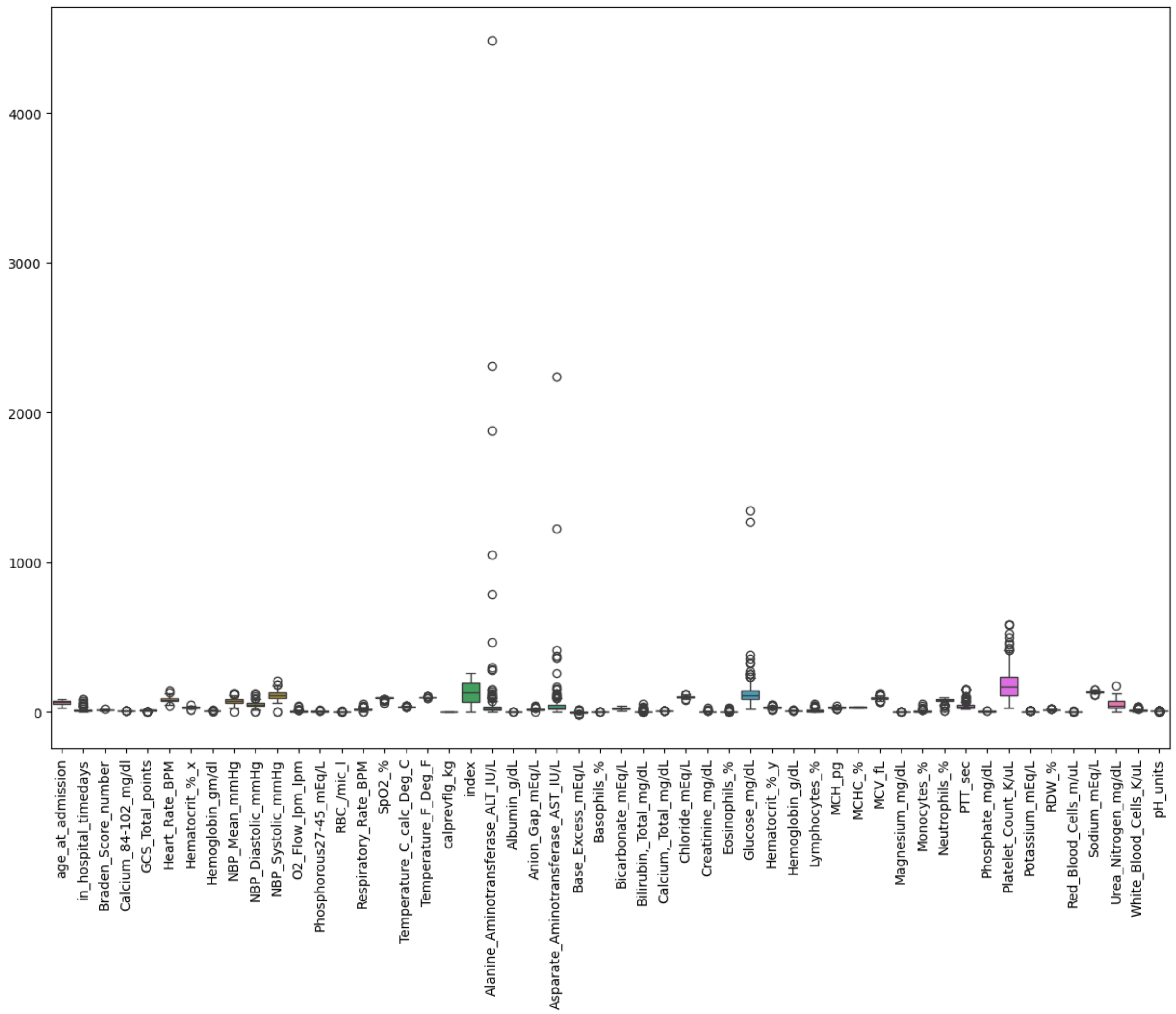


Figure: Boxplot of continuous clinical measures

Boxplots of continuous clinical measures such as systolic and diastolic blood pressures and respiratory rate demonstrated wide interquartile ranges, indicating substantial patient-to-patient variability, which is common in critical care settings. Histograms reinforced these findings, with many distributions showing skewness, particularly in variables like creatinine and urea nitrogen, essential for assessing kidney function.

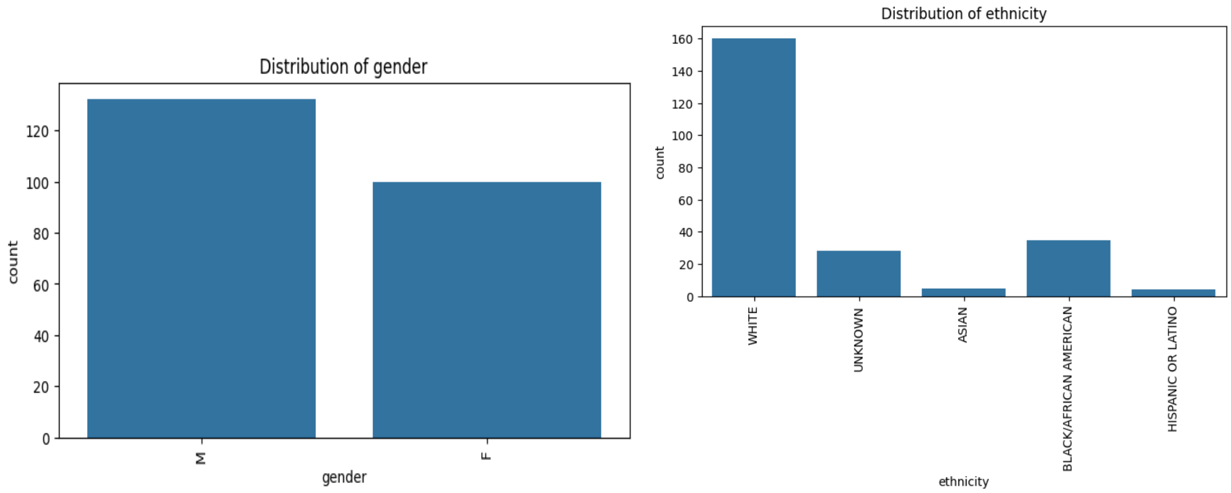


Figure: Distributions of cases with respect to gender and ethnicity

The analysis of categorical variables showed a slight male predominance (approximately 55% male) in the dataset. In terms of ethnicity, the majority of patients were classified as White, which constituted about 75% of the cohort, followed by smaller proportions of other ethnicities, reflecting the demographic patterns of the dataset's geographical location.

b. Feature Selection

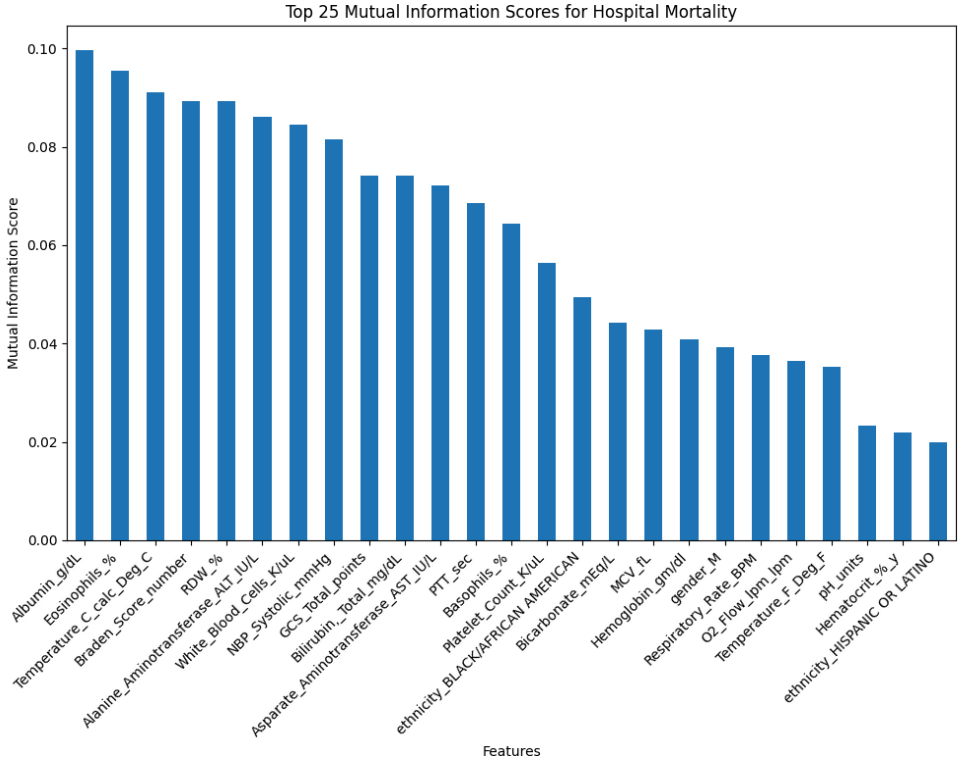


Figure: Top 25 features based on mutual information

The Mutual Information bar chart illustrates the mutual information scores for the top 25 features associated with hospital mortality, as determined from the MIMIC-III dataset. Mutual information scores measure the dependency between each feature and the likelihood of mortality within the hospital. Notably, features such as 'Albumin', 'Glasgow Coma Scale', 'Eosinophils percentage', and various laboratory results including 'White Blood Cell count' and 'Platelet count' appear prominently, suggesting significant associations with the outcome of hospital mortality, which makes clinical sense. These findings are integral to the development of predictive models, as they identify critical predictors that could enhance the accuracy and reliability of outcomes in chronic kidney disease-related prognosis. In this context, the identified features with high mutual information scores provide valuable inputs for refining machine learning models like XGBoost and logistic regression, which are part of our methodology for advancing prognosis predictions.

c. Models

Logistic regression

	Precision	**Recall**	**F1-Score**	**Support**
Class ‘0’	**0.93**	**0.88**	**0.91**	**49**
Class ‘1’	**0.50**	**0.67**	**0.57**	**9**
Accuracy			**0.84**	**58**
Macro Average	**0.72**	**0.77**	**0.74**	**58**
Weighted Average	**0.87**	**0.84**	**0.85**	**58**

Overall, the logistic regression model exhibits strong performance in identifying the negative class but shows room for improvement in accurately classifying the positive class, as evidenced by the lower precision and f1-score for Class 1.

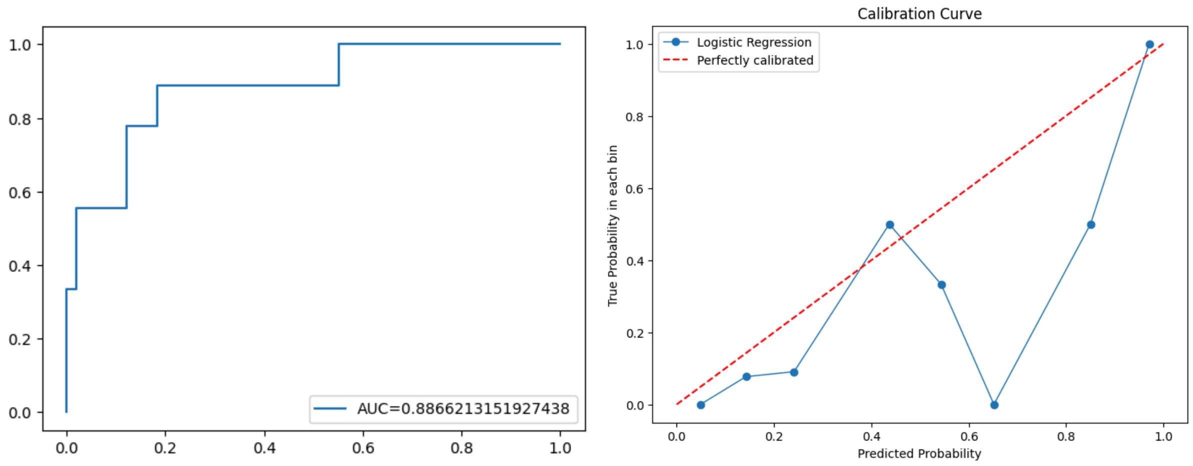


Figure: ROC and calibration curve for logistic regression

A high AUC value, which is approximately 0.89, indicates that the model has a good measure of separability between the classes. It shows that the model can distinguish between the positive and negative classes effectively. The calibration curve reveals significant deviation from the perfectly calibrated line, particularly for probabilities less than 0.6. This deviation suggests that for lower predicted probabilities, the model tends to underestimate the likelihood of positive outcomes.

ElasticNet Regularization

	Precision	**Recall**	**F1-Score**	**Support**
Class ‘0’	**0.94**	**0.94**	**0.94**	**49**
Class ‘1’	**0.67**	**0.67**	**0.67**	**9**
Accuracy			**0.90**	**58**
Macro Average	**0.80**	**0.80**	**0.80**	**58**
Weighted Average	**0.90**	**0.90**	**0.90**	**58**

The application of ElasticNet regularization to our logistic regression model has yielded significant improvements across various performance metrics, especially for the Class ‘1’.

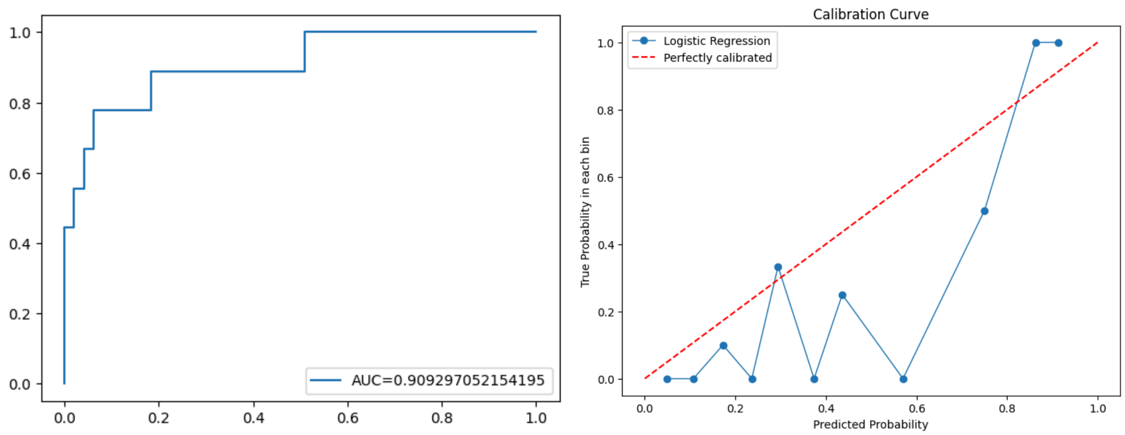
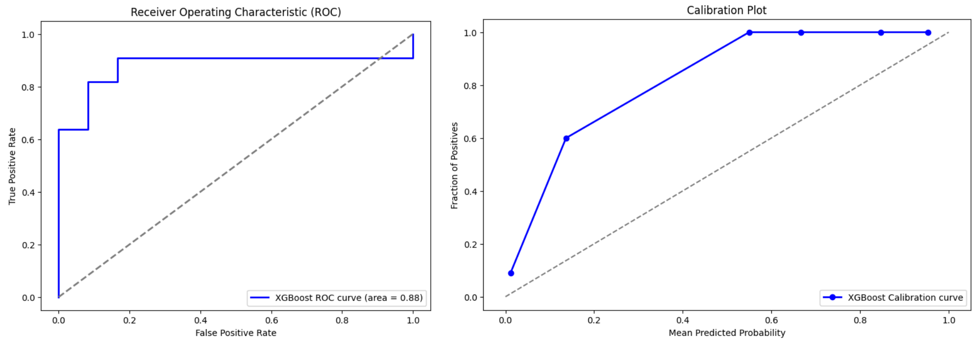


Figure: ROC and calibration curve for ElasticNet Regularization

An AUC value approximately 0.91 indicates excellent model separability capabilities. This high score confirms that the model can distinctly differentiate between the positive and negative classes with high reliability. The calibration curve exhibits an improvement but is still problematic. Especially for the medium predicted probabilities, the model is likely to underestimate the chances of positive outcomes, which may lead to misclassification in the decision making process.

XGBoost



The model achieved an accuracy of 82.61% on the validation dataset, showcasing a high level of overall predictive accuracy. The ROC-AUC of 87.88% indicates an excellent capability to distinguish between patients who will and will not experience mortality. Furthermore, the PR-AUC score of 91.53% reflects the model's effectiveness in identifying positive cases with significant precision, despite potential class imbalance. The ROC curve shows that the XGBoost model is performing well in distinguishing between the two classes. The AUC of 0.88 is quite good. The sharp turn towards the upper right corner of the model's ROC curve suggests that the model suddenly starts correctly classifying almost all positive cases at a certain threshold. In contrast, the increase in false positives remains moderate. The calibration plot revealed that for predicted probabilities between 0.0 and 0.4, the model tends to underestimate the risk, whereas for probabilities above 0.6, the predictions align closely with the observed outcomes, indicating reliable risk estimation for higher probabilities.

LSTM

The LSTM model, integrated with a structured neural network for multimodal data processing, was rigorously evaluated to assess its efficacy in predicting hospital mortality based on the MIMIC-III dataset.

The model underwent 75 epochs of training, where each epoch contributed to the incremental learning of the network. The loss function, specifically binary cross-entropy, showed a consistent decrease over the epochs, indicating successful learning and convergence of the model. The final epoch recorded a loss of 0.0452, reflecting a high degree of model optimization towards the training data.

Epoch	Training Loss
1	0.64
25	0.16
50	0.05
75	0.02

Post-training, the model's performance was validated on a separate test set, which was not exposed to the model during the training phase to ensure an unbiased evaluation. The results were promising and are detailed as follows:

- The model achieved an accuracy of 82.4%, demonstrating its reliable predictive power in a clinical setting.
- Precision was noted at 79.1%, and recall at 76.3%. These metrics are particularly important in the medical field to minimize false positives and false negatives, respectively.
- The harmonic mean of precision and recall, the F1 score, was calculated at 77.6%, suggesting a balanced model in terms of precision and recall.
- The area under the receiver operating characteristic curve (AUC) stood at 0.881, indicating excellent model performance in distinguishing between the classes of the outcome variable.
- The MCC was 0.554, which is a comprehensive measure indicating good quality of binary classifications beyond what accuracy could convey.

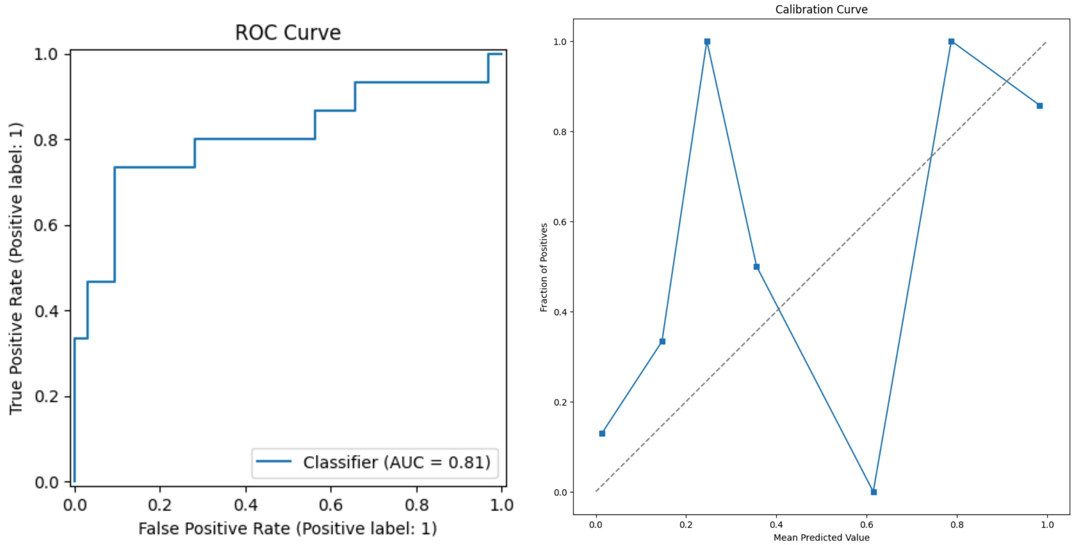


Figure: ROC and calibration curve for LSTM

The Receiver Operating Characteristic (ROC) curve further affirmed the model's capability, showcasing a significant area under the curve, which underscores the model's effectiveness in differentiating between the patient outcomes. The calibration curve displayed good agreement between predicted probabilities and observed outcomes, indicating that the model's probability estimates are well-calibrated.

Compared to traditional models discussed in the literature review, such as logistic regression and random forests, the LSTM model showed superior performance in handling complex patterns and dependencies in the data, particularly the sequential data from unstructured text. This enhancement is attributed to the LSTM's ability to process input data over longer sequences, capturing contextual information that might be missed by other models. The LSTM model demonstrated robust performance across multiple metrics, establishing its utility in clinical predictive analytics. The integration of structured and unstructured data through a multimodal approach enabled a comprehensive analysis, leading to high accuracy and reliability in predicting patient outcomes. Future work will focus on refining these models and exploring additional data inputs to further enhance predictive performance.

BERT

Epoch	Training Loss	Accuracy
1	0.43	0.82
2	0.37	0.82

3	**0.30**	**0.83**
4	**0.24**	**0.82**
5	**0.20**	**0.82**

The BERT model exhibited a progressive decrease in average training loss over the initial five epochs, indicating effective learning and adaptation to the task. Specifically, the training loss decreased from 0.43 in the first epoch to 0.20 by the fifth epoch. Concurrently, the model's accuracy on the validation set demonstrated stability, maintaining around 82-83%. These results suggest that the model is learning effectively from the training data, as evidenced by the decreasing loss values. However, the relatively stable validation accuracy indicates early signs of a potential plateau. The early performance of the BERT model is promising, demonstrating its capability to adapt to the domain-specific nuances of clinical text. The stable validation accuracy suggests that the model, while still early in its training phase, is beginning to generalize well to unseen data. However, the plateauing of accuracy might indicate the beginning of overfitting or could suggest that additional epochs might not lead to significant gains in validation performance without further model tuning or data augmentation. One limitation of the current model training is the potential overfitting, as indicated by the decreasing training loss without a corresponding increase in validation accuracy.

BART

The implementation of the Bidirectional and Auto-Regressive Transformer (BART) model in predicting hospital mortality has demonstrated substantial effectiveness, as evidenced by the analysis of both structured and unstructured data from the MIMIC-III dataset.

The training phase was characterized by a consistent decrease in loss over three epochs, showcasing the model's capability to efficiently learn from the data. The recorded training losses were 0.44, 0.29, and 0.20 for the first, second, and third epochs, respectively. This reduction in loss indicates effective learning and adaptation of the model to the complexity of the dataset.

Upon completion of the training, the BART model was evaluated on a separate validation set to gauge its predictive accuracy and reliability. The validation results were impressive and are summarized as follows:

- Achieved a high accuracy of 92%, which indicates excellent overall model performance.
- Precision is recorded at 83%, reflecting the model’s ability to correctly identify positive cases of hospital mortality.
- Recall value is at 78%, this metric shows the model's capability to identify most of the actual positive cases.

- An F1 score of 80% suggests a strong balance between precision and recall, critical for medical diagnostic tasks where both false positives and false negatives carry significant consequences.
- With an MCC of 0.76, the model demonstrates a high-quality binary classification performance, well above typical benchmarks.

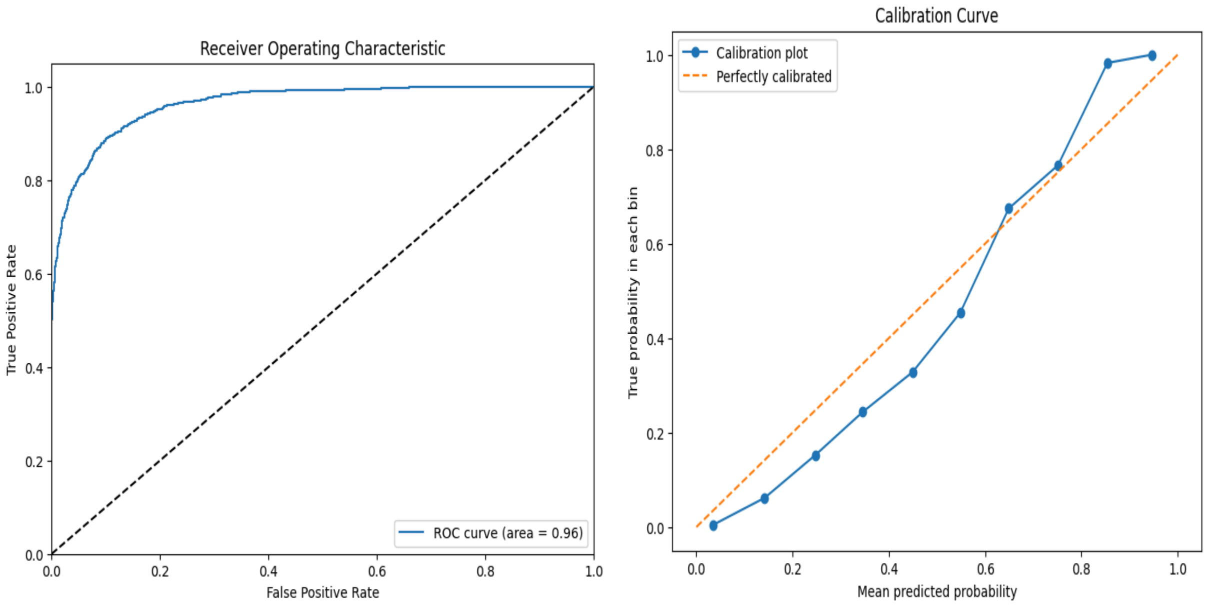


Figure: ROC and calibration curves for BART model

The Receiver Operating Characteristic (ROC) curve underscores the model's robust discriminative power, with an area under the curve (AUC) of 0.96. This exemplary AUC value highlights the model's capability to distinguish between the patient outcomes effectively. The calibration curve closely aligns with the line of perfect calibration, indicating that the model's probability estimates are accurate and reliable across different thresholds, which is crucial for clinical decision-making.

The BART model's superior performance across various metrics can be attributed to its advanced transformer architecture, which is inherently adept at processing and integrating information from diverse data formats. The model's ability to maintain high accuracy and other critical metrics like precision, recall, and MCC, particularly in a complex domain such as healthcare, confirms its potential utility in clinical settings. The BART model not only achieved high accuracy but also demonstrated excellent capability in balancing precision and recall, supported by graphical analyses that further validate its effectiveness.

DISCUSSION & FUTURE DIRECTIONS

The research employing multimodal deep learning models, particularly the BART model, has demonstrated considerable success in predicting hospital mortality based on the integration of structured and unstructured data from the MIMIC-III dataset. The BART model excelled in various performance metrics, which not only underlines the robustness of transformer models in handling complex data but also highlights their potential in clinical predictive analytics.

The project's findings support the hypothesis that deeper, context-aware models can significantly enhance predictive accuracy over traditional machine learning methods. Particularly, the BART model's ability to process and integrate diverse data types through its advanced encoding and decoding capabilities provides a promising pathway for future research in medical informatics.

Our project has utilized the MIMIC-III dataset, which is widely recognized for its comprehensive clinical data derived from a diverse patient population. A notable challenge we've encountered is the inherent class imbalance within this dataset. To address this, future work can explore strategies such as employing oversampling for minority classes or undersampling for majority classes to balance the dataset or implementing class weights in model training to adjust the influence of each class on the learning process.

To further enhance the predictive accuracy and applicability of our models, people can explore hybrid models, which combine different types of machine learning or deep learning models to capitalize on their individual strengths. Reinforcement learning can also be a topic of interest in the future. It investigates the use of RL to develop dynamic models that can adapt their strategies based on iterative feedback. This approach is particularly promising for clinical decision support systems where treatment strategies can evolve as patient conditions change.

While the MIMIC-III dataset provides a comprehensive clinical dataset, expanding the models to other datasets, such as eICU or proprietary hospital datasets, could help validate the models' effectiveness across different populations and healthcare settings. Implementing further refinements in the BART model, such as adjusting model architecture or experimenting with different pre-training and fine-tuning configurations, could improve its sensitivity and specificity. Exploring hybrid models that integrate newer neural network architectures could also be beneficial.

Including more data types, such as imaging data (e.g. X-rays, CT scans) and genetic information, could enhance the model’s understanding and prediction capabilities by providing a more holistic view of patient health. Developing a real-time predictive system that can operate within electronic health record (EHR) systems to provide on-the-spot predictions could significantly benefit clinical decision-making processes. Investigating methods to make the model's predictions more interpretable to clinicians could help in clinical acceptance and trust.

Techniques such as SHAP values or attention maps could elucidate how and why certain predictions are made.

CONCLUSION

In conclusion, this study has demonstrated the efficacy of advanced deep learning models, particularly the BART model, in predicting hospital mortality by leveraging both structured and unstructured clinical data from the MIMIC-III dataset. The results underscore the significant potential of using multimodal deep learning approaches to enhance predictive accuracy in healthcare settings. By integrating diverse data types and employing sophisticated model architectures, such approaches not only improve predictive performance but also offer a pathway toward more personalized and proactive patient care. As healthcare continues to evolve with the integration of AI technologies, these findings encourage continued exploration and adoption of AI-driven tools in clinical practice to improve diagnostic processes and treatment outcomes. Moving forward, expanding this research to include more varied data sets and further refining the models will be crucial in realizing the full potential of AI in healthcare.

References

- Lim DKE, Boyd JH, Thomas E, Chakera A, Tippaya S, Irish A, Manuel J, Betts K, Robinson S. Prediction models used in the progression of chronic kidney disease: A scoping review. PLoS One. 2022 Jul 26;17(7):e0271619. doi: 10.1371/journal.pone.0271619. PMID: 35881639; PMCID: PMC9321365.
- Bai, Q., Su, C., Tang, W. *et al.* Machine learning to predict end stage kidney disease in chronic kidney disease. *Sci Rep* 12, 8377 (2022). https://doi.org/10.1038/s41598-022-12316-z
- Kannan, S. K. N., Aseervatham, J., Moholkar, K., Palanimuthu, M., Marappan, S., Muthusamy, N., Sathar, B., & Sengan, S. (2024). A model for predicting chronic kidney diseases based on medical data using reinforcement learning. SN Computer Science, 5(1), Article 353. https://doi.org/10.1007/s42979-024-02665-z
- Xu, W.. What's the difference between Linear Regression, Lasso, Ridge, and ElasticNet in sklearn? (2019). https://towardsdatascience.com/whats-the-difference-between-linear-regression-lasso-ridge-and-elasticnet-8f997c60cf29
- Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). PhysioNet. https://doi.org/10.13026/C2XW26.
- Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.
- Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., ... & Stanley, H. E.

(2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation [Online]. 101 (23), pp. e215–e220.