Towards Interpretable Multimodal Predictive Models for Early Mortality Prediction of Hemorrhagic Stroke Patients

Forhan Bin Emdad, MS[1], Shubo Tian, PhD[1], Esha Nandy, MS[1], Karim Hanna, MD[2], Zhe He, PhD[1]

[1]Florida State University, Tallahassee, Florida, USA

[2]Morsani College of Medicine, University of South Florida, Tampa, Florida, USA

Abstract

The increasing death rate over the past eight years due to stroke has prompted clinicians to look for data-driven decision aids. Recently, deep-learning-based prediction models trained with fine-grained electronic health record (EHR) data have shown superior promise for health outcome prediction. However, the use of EHR-based deep learning models for hemorrhagic stroke outcome prediction has not been extensively explored. This paper proposes an ensemble deep learning framework to predict early mortality among ICU patients with hemorrhagic stroke. The proposed ensemble model achieved an accuracy of 83%, which was higher than the fusion model and other baseline models (logistic regression, decision tree, random forest, and XGBoost). Moreover, we used SHAP values for interpretation of the ensemble model to identify important features for the prediction. In addition, this paper follows the MINIMAR (MINimum Information for Medical AI Reporting) standard, presenting an important step towards building trust among the AI system and clinicians.

Introduction

Hemorrhagic stroke is one of the leading causes of death and a major cause of disability in the United States according to the trend report by the Centers for Disease Control and Prevention (CDC).[1] Hemorrhagic stroke occurs due to bleeding into the brain which is caused by rapture of the blood vessel. Hemorrhagic stroke is further divided into intracerebral hemorrhage (ICH) and subarachnoid hemorrhage (SAH). According to a recent survey, 35% of stroke patients die within 7 days of the stroke and about 50% of intracerebral hemorrhagic stroke patients died within 30 days.[2] Most hemorrhagic stroke patients are admitted to intensive care units (ICUs) after stroke.[3] Early prediction of mortality and identification of factors related to the mortality of hemorrhagic stroke patients in the ICU setting can potentially reduce mortality rate through targeted interventions.

Currently, most clinicians use traditional risk scores, as well as simple statistical and machine learning (ML) models for mortality prediction.[4] Several risk scores such as Acute Physiology and Chronic Health Evaluation (APACHE II, IV), Simplified Acute Physiology Score III (SAPS III) have been evaluated for predicting mortality of stroke patients.[8] Simple statistical and ML algorithms like logistic regression, decision tree, and random forest were also used for mortality prediction. Still many clinicians rely on the risk scores due to their simple and understandable structure even though these scores consider very limited number of features with suboptimal sensitivity and specificity.

In recent years, the advent of machine learning (ML) and deep learning (DL) has significantly improved the accuracy of predictive analysis in healthcare.[5] The increasing use of electronic health record systems (EHRs) in hospitals and other clinical settings has made it possible to develop more advanced ML and DL models for mortality predictions.[6] As such, many researchers have used EHR data to build prediction models with ML and DL approaches in the healthcare.[7] Specifically, deep learning models have been developed for predicting ICU mortality with significant accuracy. However, use of AI in healthcare faces significant challenges such as following practice standards, lack of interpretation and unexpected low performance.[9] Furthermore, we have not well explored how to leverage data of different modalities such as snapshot, time series, textual, image, and audio data for mortality prediction by ML and DL models. Addressing these challenges can build a bridge of trust among the health informatics researchers, clinicians, physicians, and stakeholders.

While much research has studied use of ML and DL models for prediction of different mortalities, very few studies have been conducted to explore ML and DL for stroke mortality prediction. In this study, we extracted a cohort of stroke patients from the MIMIC-III database and developed a dataset for experiment of using ML and DL algorithms for mortality prediction of stroke patients. Following the MINimum Information for Medical AI Reporting (MINIMAR) standard, we developed an ensemble model to predict stroke patient mortalities based on their aggregated and hourly measured data. Our ensemble model was able to achieve significantly better performance compared to traditional machine learning algorithms. In addition, we used SHAP values for interpretation of our ensemble model to identify the important features for prediction. In this context, this study has the following overarching goals:

1) Designing interpretable multimodal predictive models for hemorrhagic stroke mortality prediction following the standard medical AI reporting guidelines.

2) Comparing the performance of the designed models with the baseline models.

3) Providing explanation and interpretation of the models.

Related Work

There were some interesting studies which predicted stroke patients' mortality with simple models. One of those studies is Nie's study where traditional models were used with mean, standard deviation, and maximum, minimum of variables for predicting 7 days and 30 days of intracerebral hemorrhage stroke patient mortality.[3] Although Nie's study was related to mortality prediction in cerebral hemorrhage stroke patients, the traditional models (nearest neighbors, decision tree, neural net, Adaboost, random forest) used in the study were not able to produce more than 70% of accuracy. Similarly, Scrutinio's random forest model achieved highest accuracy of 77% for mortality prediction of stroke patients.[11] To improve the prediction performance from the previous studies, we identified few studies which implemented multimodal algorithm approach for mortality prediction but these studies were not specifically focused on hemorrhagic stroke patients' mortality prediction. Purushothom's benchmarking models are considered to be one of the best models for EHRs data analysis and mortality prediction.[10] In Purushothom's study, multimodal DL (MMDL) achieved the highest area under the ROC curve (AUROC) score of 92% in mortality prediction of 1-day, 2-day, 30-day, and 1-year using MIMIC data. However, Purushothom's study was about in-hospital mortality prediction instead of hemorrhage patient mortality prediction and the selected features in the study were limited. In another study, multitask learning with MIMIC time series data was performed for predicting hospital mortality.[12] Zhang's study shed light on the use of fusion techniques using CNN and LSTM.[13] Both unstructured and structured data were used in Zhang's study. Similarly, Xu's paper also provided multimodal fusion architecture (MUFASA) for diagnosis using EHR data and the performance of the designed model outperformed Transformer based models.[14] Similar to our techniques, previous studies proposed multimodal fusion and ensemble techniques for mortality prediction but struggled to achieve higher performance in terms of accuracy, and provide a proper explanation of the models.

Method

Figure 1 provides the overall representation of the workflow of this study, consisting of dataset development, feature selection, preprocessing, modeling with ML and DL, and interpretation with SHAP values following the MINIMAR standard. In the remainder of the Method section, we will describe the details of each step of this workflow.

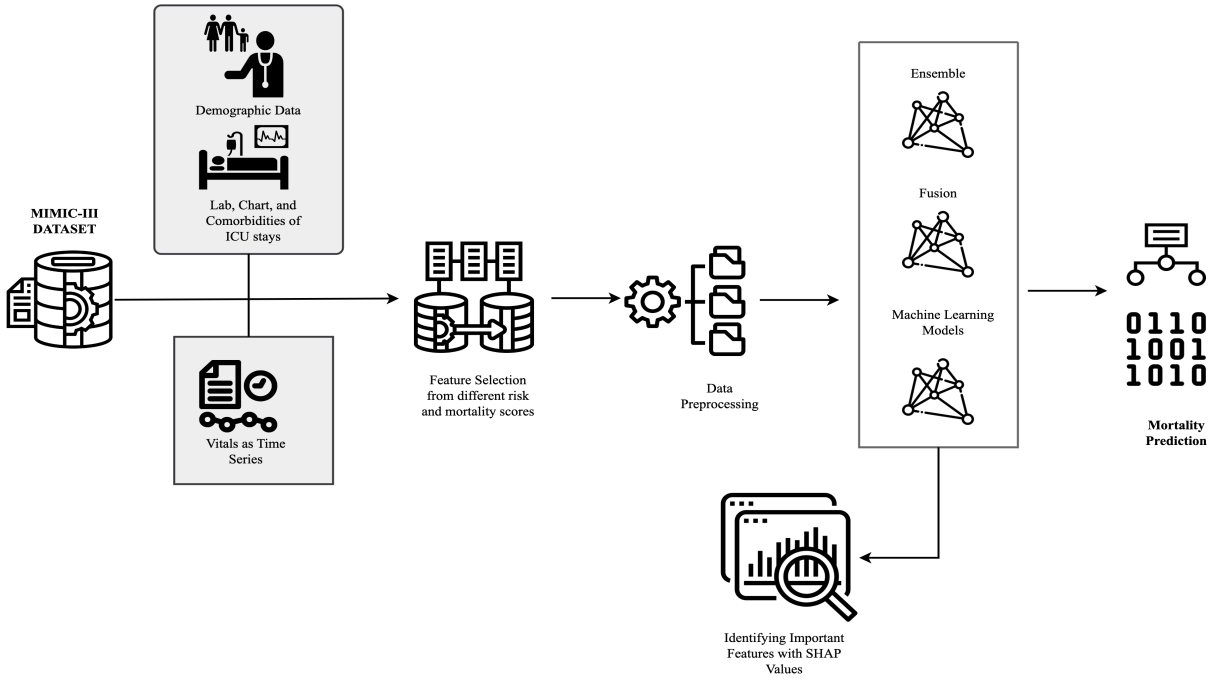

Figure 1. The workflow of the study

MINIMAR

Adhering to a reporting guideline during the design of AI systems will make the AI more trustworthy to the stakeholders. The MINIMAR (MINimum Information for Medical AI Reporting), which consists of 4 components: study population and setting, patient demographic characteristics, model architecture, and model evaluation, is a popular reporting standard for artificial intelligence in healthcare.[22] We followed the guidelines provided by the MINIMAR standard in this study.

Data Source

We used data extracted from MIMIC-III (Medical Information Mart for Intensive Care) database. MIMIC-III is a clinical database that contains data on patients hospitalized to a large hospital's critical care units.[15] During the data collecting phase of MIMIC-III, two critical care information systems were used: Philips CareVue Clinical Information System and iMDsoft MetaVision ICU. Vital signs, medications, laboratory measures, care providers' observations and notes, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more are all included in the MIMIC-III database. MIMIC-III provides data on 53,423 unique hospital admissions for adult patients (aged 16 and over) admitted to critical care units between 2001 and 2012. MIMIC-III database is a relational database which contains 26 tables. Among the tables, we have included 8 tables (ADMISSIONS, CHARTEVENTS, DIAGNOSES_ICD, ICUSTAYS, OUTPATIENTS, LABEVENTS, PATIENTS, SERVICES) as we are considering vital signs and laboratory measures, social determinants, interventions, and demographics to predict ICU mortality within 7 days.

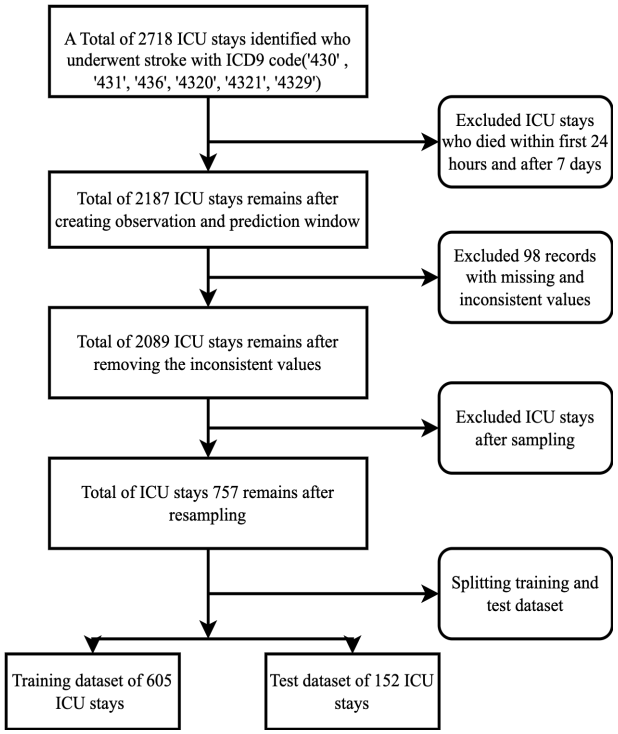

Figure 2. The detailed process of data extraction using exclusion-inclusion criteria.

Cohort Definition

As shown in Figure 2, we extracted unique ICU stays by constraining with the ICD-9-CM codes related to hemorrhagic stroke ('430', '431', '436', '4320', '4321', '4329'). This cohort consisted of 2,718 ICU stays. Later, another constraint was applied by excluding ICU stays who died within 24 hours. Applying the constraint, the total number of ICU stays became 2,187. We then removed 98 ICU stay records with inconsistent and missing values. The final total number of ICU stays remains 2,089 (mortality rate = 14.0%).

Outcome Definition

As illustrated in Figure 3, we define observation window as the first 24 hours and prediction window as the subsequent six days. In the observation window, we collected ICU patients' vital signs and other relevant factors. In the prediction window, we captured the mortality information as the outcome.

Feature Selection

To identify relevant features, we considered different popular risk scores, mortality scores, and organ dysfunction scores such as OASIS, APS II, SAPS, SAPS II, APACHE II, APACHE IV, LODS, qSOFA, SOFA, and SIRS.[15-17] We used 73 variables consisting of demographics, lab and chart values, comorbidities, vitals, and other variables. In other variable types, we included temperature, glucose, Glasgow Coma Scale (GCS) score for motor (gcsmotor), verbal (gcsverbal), and eyes (gcseyes), endotracheal position, urine output, services related to orthopedic medicine (service_omed), neurologic medicine (service_nmed), neurologic surgery (surgery_intervention), thoracic surgery (service_tsurg), cardiac surgery (service_cmed), vascular surgery (service_vsurg), plastic surgery (service_psurg), urinary system (service_gu), general surgery (service_surg), female reproductive systems (service_gyn), and trauma (service_traum). The list of variables is given in Table 1.

Furthermore, categorical variables such as gender and race were converted into numerical variables. We categorized the variables into two categories: temporal, and aggregated data. Temporal variables consisted of vital signs. Aggregated data consisted of all other variables.

In this research, hourly data for vital signs were included in the observation window. As other relevant features did not have hourly values, we employed the statistical approach of taking the minimum, mean, standard deviation, and maximum value of other variables as potential features.

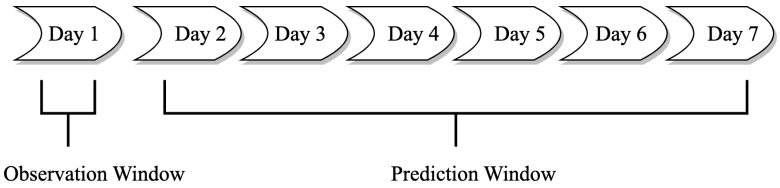

Figure 3. Visualization of the observation and prediction window

Table 1. Details of variable types and variables

Variable Type	Variables
Demographics	Age, Gender, Race
Lab & Chart Value	PO2 (partial pressure of oxygen), partial pressure of carbon dioxide (PCO2), the ratio of arterial oxygen partial pressure (PaO2 in mmHg) to fractional inspired oxygen (FiO2 expressed as a fraction, not a percentage), measure of the acidity or alkalinity, presence of strong acid (metabolic acidosis) or strong base (metabolic alkalosis), level of bicarbonate, combination of hemoglobin and carbon monoxide formed in the blood, hemoglobin found in the blood in small amounts, aniongap, albumin bands, bicarbonate, bilirubin, calcium, creatinine, chloride, hematocrit, hemoglobin, lactate, platelet, potassium, Partial thromboplastin time (PTT), international normalized ratio from prothrombin time (PT), sodium, blood urea nitrogen (bun), white blood cells count (wbc)
Comorbidities	heart failure, hypertension, metastatic cancer, obesity, alcohol abuse, depression, paralysis, diabetes, weight loss, drug abuse
Vitals	heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, respiratory rate, oxygen saturation (SpO2)

Class Imbalance Correction

Extracted data were imbalanced as the number of survivors was 1797 and the number of non-survivors was only 292. Class imbalance is a common and significant problem for predictive modeling using EHR data.[20] Sampling is one of the solutions for tackling imbalanced classification problem. Imputation techniques vary depending on the type of the study. Synthetic minority over-sampling (SMOTE),[21] down/under sampling, or up sampling are some of the popular sampling techniques for machine learning classification tasks. We used under-sampling technique which randomly removes samples from the majority class (negative samples) to balance with the minority class. After under-sampling the negative instances, the total number of ICU stays became 757 pertaining to 742 patients. Among them, 450 were surviving ICU stays and 307 were non-surviving ICU stays. The mortality rate became 40.6%.

Data Imputation

Data quality issue is a major challenge for EHR data. EHR data contains a lot of missing values and inconsistent data.[18] To address this challenge, different imputation techniques have been used in EHR-based predictive modeling. Mean and median imputation are the most popular imputation techniques which can be performed quite easily. In addition, there are other imputation techniques such as K-nearest neighbors based (KNN) imputation, MissForest, and multiple imputation by chained equations (MICE). MICE imputation technique is proven to generate better performance with EHR data.[19] Hence, we chose to use MICE imputation in our study.

Machine Learning and Deep Learning Algorithms

We will introduce various machine learning and deep learning algorithms including ensemble, fusion, and baseline models used in this study in the subsections below.

Fusion Model: Fusion is a technique of creating a better knowledge representation by concatenating data from multiple modalities aiming to achieve better ML performance than single modality.[27] In case of our study, we gathered vital signs data as time series data and non-vital sign data as aggregated data for the fusion model.[16] In this context, our data can be considered a multimodal data. Various fusion architectures (early, late, and joint) can be applied to the fusion model.[28] Early fusion technique is joining the data at the initial input level before feeding them to the neural network. Late fusion is the process where the fusion takes place at the prediction level to generate a final prediction. Joint fusion is the process of concatenating learned knowledge representations from multiple models and then feed the concatenated knowledge representation to another model as input to generate final prediction. We used joint fusion mechanism for our fusion model. In this study, time series data have learned from 2-LSTM layers (128 and 64 neurons in each layer, respectively) and aggregated data have learned from 2-fully connected layer (128 and 64 neurons in each layer, respectively). Later, the combined output was fed to a fully connected layer to generate the final prediction output of mortality. Long short-term memory (LSTM) is a specific kind of recurrent neural network that contain three multiplicative units which are the input, output and the forget gates that provide continuous analogues of write, read, and reset operations for the cells.[26] Figure 4 shows the model architecture of the fusion model.

Despite proper data preprocessing, optimization of the algorithm is necessary for building an optimal model. We experimented with different numbers of neurons, layers, activation function, optimizers, batch size, and epochs in our models and achieved the best results with the lowest loss rate, highest and faster learning rate, and highest accuracy. We tried different combinations of the hyper parameters and selected the best combination. We tried batch size "16", "32", and "64" to find the best batch size and chose "32" as the batch size. Similarly, we tried "20", "50", and "100" epochs for the model, later found "100" epochs to be the best. We also tried different optimizers such as "SGD" and "RMSProp", but found that the most effective optimizer is the Adam optimizer which is an algorithm for first order gradient-based optimization of stochastic objective function.[29] This optimizer is efficient for handling a large number of data and parameters. Adam optimizer is often used for models using EHR data.[30] As our data was highly imbalanced, we used focal loss error in the fusion model.[31] Like object detection, focal loss corrects class imbalance during training. For activation functions, 2-LSTM layers and 2 fully connected layers use a piecewise linear function (ReLU) which passes the output as positive or zero and the final output layer uses the Sigmoid function so that the output always remains between 0 and 1 as this is a classification task.

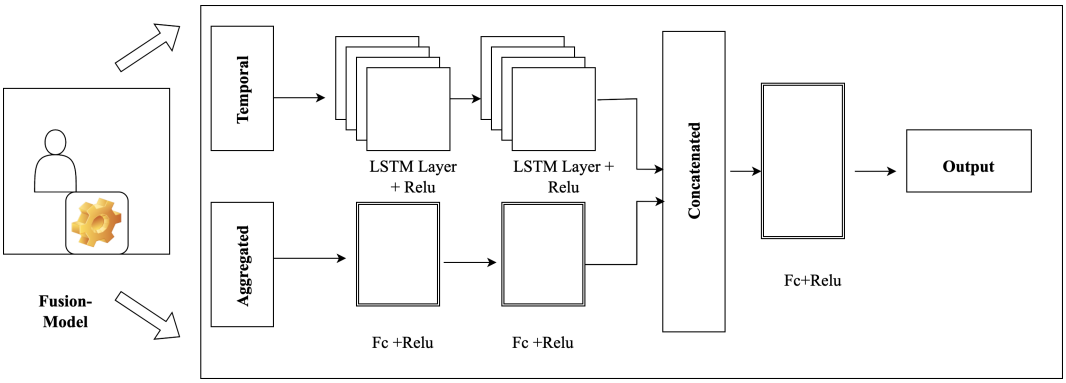

Figure 4. The model architecture of Fusion model. Fusion model uses 2-layer LSTM and 2- dropout (10%) to model temporal data. Consequently, 2-fully connected layer and 1- dropout (10%) are used to model aggregated data. Then, the output is concatenated and passed through another fully connected layer and the output layer to make predictions. Fusion model used total of 345,025 parameters.

Ensemble: The ensemble is a technique of creating a new model by combining two or more models. There are various kinds of ensemble methods such as averaging, max voting, stacking, blending, bagging, and boosting.[32] Our intention in this study is to build the ensemble model using the same models and data from fusion (before the fully connected layer). However, we used late fusion in the ensemble model. We used averaging ensemble method consisting of the two independent models and providing the average of the prediction of the two models. Binary cross entropy loss was used for loss function in both models of ensemble. In the first model (LSTM), total of 119,105 were used and second

model (fully connected) used total of 29,057 trainable parameters. Figure 5 shows the high-level overview of the ensemble model.

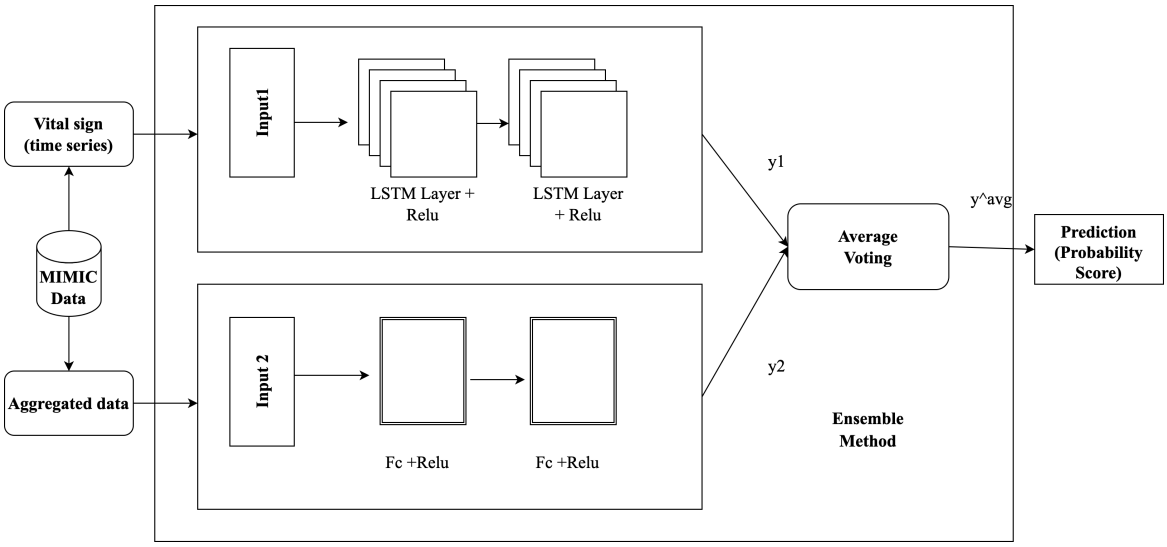

Figure 5. The high-level overview and model architecture of Ensemble average method. Like the Fusion model, Ensemble average uses 2-layer LSTM and 2- dropout (10%) to model temporal data and 2-fully connected layer and 1- dropout (10%) are used to model aggregated data. Then, predictions from both models are averaged to generate final output prediction.

Baseline Machine Learning Models. We used decision tree[23], random forest[25], XGBoost, and logistic regression[24] as the baseline model for this research. In addition, XGBoost uses extreme gradient boosting, which is easy to scale and visualize with available libraries. We converted all the features including the vital signs into summary statistics (min, max, standard deviation, mean) and used these summary statistics for each variable in the baseline models. We compared the fusion model and ensemble model performance with the baseline models for the validity of the model.

Model Evaluation

We reported Precision, Recall, F1, Accuracy and AUROC for performance evaluation of the models. Accuracy is the ratio of accurately predicted observations for patient mortality to all observations. Similarly, the ratio of accurately predicted positive (mortality) observations to total predicted positive (mortality) observations is known as precision. On the other hand, recall is defined as the proportion of accurately predicted positive (mortality) observations to all the positive observations in the class. F1 is the harmonious mean of precision and recall. The area under the ROC curve (AUROC) estimates the capability of Fusion and Ensemble models to differentiate between survival and non-survival. We also performed 5–fold cross validation on the baseline ML models to validate the generalizability of these models. We evaluated the averages of all the performance metrics for 5–fold cross validation and calculated standard deviations for them. However, we did not perform 5-fold cross validation for the fusion and ensemble models as they contain data with different modalities.

Explanation of the Models with SHAP

We used SHAP values (SHapley Additive exPlanations) for providing transparency and interpretability of the ensemble model. SHAP values is the process of assigning value to the features which reflects the relations of the features with the output. This explainability method was derived from coalitional game theory to identify an approach to disseminate the "pay" to all features properly. "shap.DeepExplainer" package was used for deriving the SHAP values which uses the DeepLIFT algorithm (Deep SHAP). In addition, we used "shap.summary_plot" package to provide visualization of the important features.

Results

Our study has checked almost all the feature components of MINIMAR reporting guideline except internal and external model validation. Providing standards to our report will help the medical informatics research community to evaluate our models with ease.

Baseline Characteristics

As shown in Table 2, most of the ICU stays in the cohort involved patients above 70 years, and majority race was white. Although, the male and female percentage of the cohort were almost equal.

Table 2. Demographic characteristics for ICU stays

	Positive Instances		**Negative Instances**	
Age	**Number**	**Mean (Standard Deviation)**	**Number**	**Mean (Standard Deviation)**
0-19	0	0	0	0
20-45	21	36.6 (6.29)	62	37.9 (5.63)
45-65	87	57.16 (5.59)	169	57.02 (5.12)
65-75	67	70.53 (3.32)	90	71.78 (2.68)
75 and above	128 (excluding 4 inconsistent age values (age > 300))	82.06 (3.1)	124 (excluding 4 inconsistent age values (age > 300))	81.5 (3.43)
Gender	**Number**	**Percentage**	**Number**	**Percentage**
Male	158	51.47	244	54.22
Female	149	48.53	206	45.78
Race				
While	198	64.4	321	71.33
Black	18	5.9	38	8.44
Hispanic	15	4.9	13	2.9
Asian	15	4.9	11	2.44
Other Race	61	19.9	67	14.89

Performance Comparison and Analysis

Table 3 provides the performance metrics precision, recall, F1, accuracy, and AUROC for the tested models. The ensemble average model has outperformed other models in all evaluations except for AUROC. Logistic regression, random forest, and XGBoost had the highest AUROC. We can observe from comparing the results that the multimodal technique of combining temporal and aggregated features significantly increases the performance. Figure 6 gives the ROC curves of the models.

Table 3. Ensemble average (avg) and Fusion model performance comparison with the baseline models

Models	**Precision (STD)**	**Recall (STD)**	**F1 (STD)**	**Accuracy (STD)**	**AUROC (STD)**
Fusion model	0.79	0.61	0.69	0.78	0.75
Ensemble (avg)	**0.8**	**0.77**	**0.79**	**0.83**	0.82
Logistic Regression	0.77 (0.04)	0.71 (0.05)	0.74(0.04)	0.8 (0.03)	**0.87 (0.03)**
Decision Tree	0.72 (.06)	0.68 (.06)	0.7 (0.04)	0.76 (0.03)	0.80(.05)
Random Forest	0.76 (0.05)	0.74 (0.06)	0.75 (0.05)	0.8 (0.04)	**0.87 (0.03)**
XGBoost	0.75 (0.06)	0.74 (0.05)	0.75 (0.05)	0.8 (0.04)	**0.87 (0.03)**

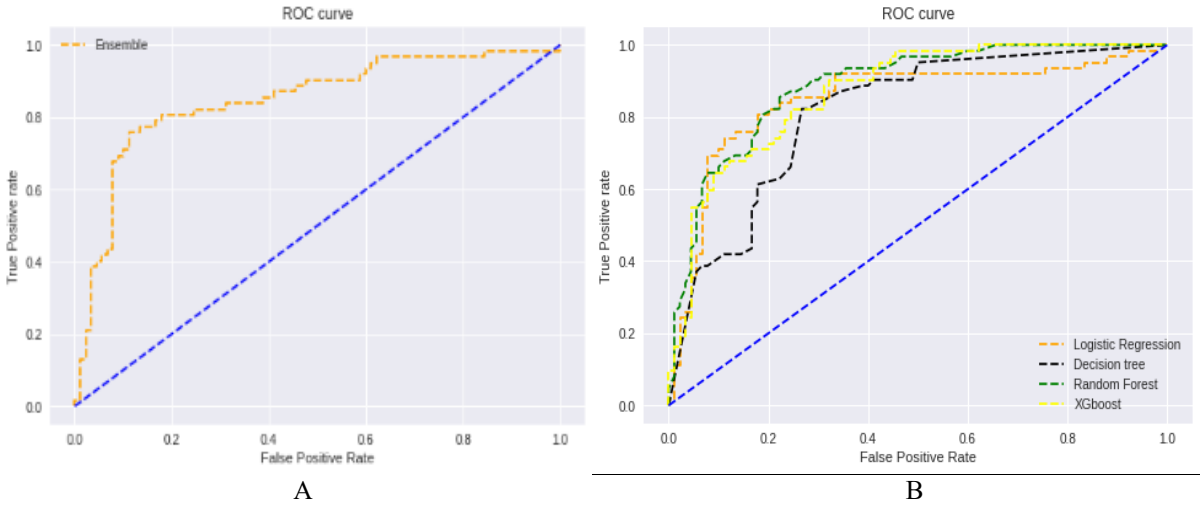

Figure 6. The ROC curves of Ensemble model and other baseline models. A) Shows ROC curve of Ensemble model. B) Shows the ROC curve of baseline models (Logistic Regression, Decision Tree, Random Forest, XGBoost).

Explainability with SHAP

As there are two models in our ensemble model, we are considering aggregated features for interpretability because temporal features consist of only 7 features. Global and local are two important methods of providing interpretability while using SHAP values.[33] Global interpretability provides the collective SHAP values to display the impact of the features on mortality prediction. On the contrary, local interpretability shows the SHAP values of each observation. In case of global interpretation of the result in Figure 7, we can find that the most important feature for mortality prediction from the aggregated features is the maximum value of GCS motor response which measures patient's ability to follow the command and move with application of different pain stimulus. Likewise, other important variables included glucose, blood urea nitrogen, overall GCS score, white blood cells count, temperature, GCS eyes response, heart failure, services related to surgery (general but not classified) and gynecology, race, and neurologic (related to brain) surgical interventions.

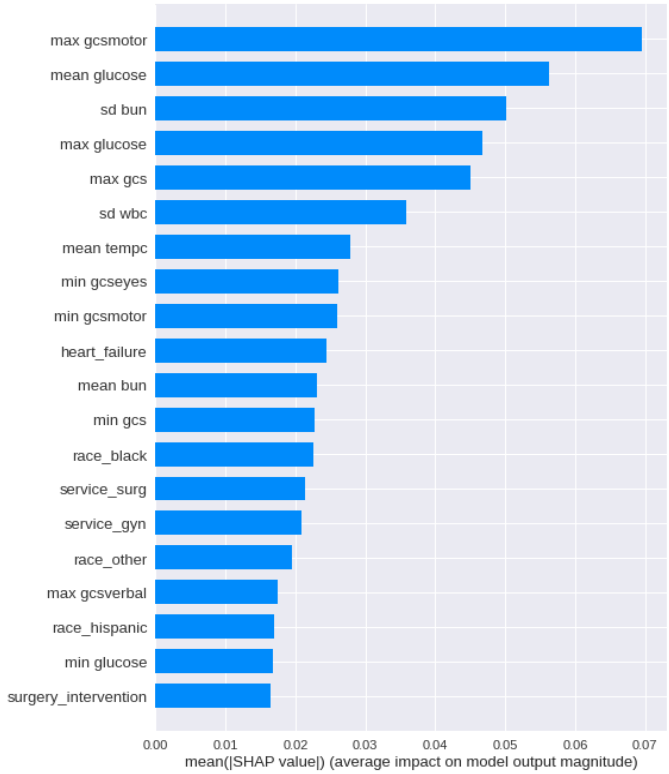

Figure 7. The summary plot shows the global interpretation of the most important features and the magnitude of the impact of the features on the Ensemble model.

Discussion

While the evaluation of patients with hemorrhagic stroke is coupled with significant burden of disease, admission for intensive care allows clinicians to stabilize patients and allows them to start a journey on the road towards recovery.[34] Without a doubt, clinical significance can be found in improved predictive models as they can guide decision making when signs, symptoms and data flag poor outcomes. This study showed that the ensemble model produces a better accuracy, precision, recall, and F1 score than the fusion model and baseline models. This guides to the understanding that vital signs collected with regular time stamps would be necessary for reducing the mortality of hemorrhagic stroke patients. This may seem to be common practice in most ICUs, but we

find that vital signs and other features of patient status should be handled in separate models for better prediction outcomes. Furthermore, other subjective and objective findings like GCS scores, labs like serum glucose, blood urea nitrogen, white blood cells count, and presence of other comorbidities like heart failure are all significant to mortality outcomes in the ICU.[35] Providers caring for patients in this acute and subacute setting should be made aware and have access to the best predictors to improve their approach to therapeutic interventions, prioritizing those most impactful in leading to morbidity and mortality. Further studies should be sought out to assess if addressing these findings would improve patient outcomes.

Looking at the limitations, racial bias was observed in the dataset, external validation of the model was not performed in this study due to lack of external data sources. Future research can address the bias issue by adopting more fair models. Moreover, data with different modalities such as neuroimaging and clinical notes can be used for mortality prediction in the future. In addition, some features such as Glasgow Coma Scale (GCS) score for motor are subjective which may have introduced human variability.

Conclusions

In this study, we presented robust multimodal deep learning predictive models combining both temporal and aggregated features from EHR data. We generated global interpretation to focus on the imported features extracted from the ensemble model. Identifying important features for mortality prediction can play a vital role in taking early precautions in clinical settings. In addition, we followed the MINIMAR reporting guidelines in our study. Further research can be conducted on this study to overcome the indicated limitations.

Data and Code Availability

The datasets used in this study can be found at: https://archive.physionet.org/works/MIMICIIIClinicalDatabase/files/. We used the Google Colab notebooks platform to implement our algorithm framework. All the predictive models of this study are available from the website at: https://github.com/ForhanBinEmdad/papers

Acknowledgments

This study was supported in part by the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Number R21AG061431 (ZH); and the University of Florida Clinical and Translational Science Institute, which is supported in part by the NIH National Center for Advancing Translational Sciences under award number UL1TR001427.

References

1. CDC. Stroke facts [Internet]. Centers for Disease Control and Prevention. 2022 [cited 2022 Aug 31]. Available from: https://www.cdc.gov/stroke/facts.htm
2. Van Asch CJ, Luitse MJ, Rinkel GJ, van der Tweel I, Algra A, Klijn CJ. Incidence, case fatality, and functional outcome of intracerebral haemorrhage over time, according to age, sex, and ethnic origin: a systematic review and meta-analysis. The Lancet Neurology. 2010 Feb 1;9(2):167-76.
3. Nie X, Cai Y, Liu J, Liu X, Zhao J, Yang Z, Wen M, Liu L. Mortality prediction in cerebral hemorrhage patients using machine learning algorithms in intensive care units. Frontiers in neurology. 2021 Jan 20;11:610531.
4. Parikh RB, Kakad M, Bates DW. Integrating predictive analytics into high-value care: the dawn of precision delivery. Jama. 2016 Feb 16;315(7):651-2.
5. Mastoli MM, Pol UR, Patil RD. Machine learning classification algorithms for predictive analysis in healthcare. Mach. Learn. 2019 Dec;6(12):1225-9.
6. International Organization for Standardization. Health informatics-Electronic health record-Definition, scope and context. na; 2005.
7. Milenkovic MJ, Vukmirovic A, Milenkovic D. Big data analytics in the health sector: challenges and potentials. Management: Journal of Sustainable Business and Management Solutions in Emerging Economies. 2019 Jan 10;24(1):23-33.
8. Johnson AE, Stone DJ, Celi LA, Pollard TJ. The MIMIC Code Repository: enabling reproducibility in critical care research. Journal of the American Medical Informatics Association. 2018 Jan;25(1):32-9.
9. Ross C, Swetlitz I. IBM's Watson supercomputer recommended 'unsafe and incorrect'cancer treatments, internal documents show. Stat. 2018 Jul 25;25.
10. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. Journal of biomedical informatics. 2018 Jul 1;83:112-34.

11. Scrutinio D, Ricciardi C, Donisi L, Losavio E, Battista P, Guida P, Cesarelli M, Pagano G, D'Addio G. Machine learning to predict mortality after rehabilitation among patients with severe stroke. Scientific reports. 2020 Nov 18;10(1):1-0.
12. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Scientific data. 2019 Jun 17;6(1):1-8.
13. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC medical informatics and decision making. 2020 Dec;20(1):1-1.
14. Xu Z, So DR, Dai AM. Mufasa: Multimodal fusion architecture search for electronic health records. InProceedings of the AAAI Conference on Artificial Intelligence 2021 May 18 (Vol. 35, No. 12, pp. 10532-10540).
15. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Scientific data. 2016 May 24;3(1):1-9.
16. Tang H, Jin Z, Deng J, She Y, Zhong Y, Sun W, Ren Y, Cao N, Chen C. Development and validation of a deep learning model to predict the survival of patients in ICU. Journal of the American Medical Informatics Association. 2022 Sep;29(9):1567-76.
17. Norman SB, Hami Cissell S, Means-Christensen AJ, Stein MB. Development and validation of an overall anxiety severity and impairment scale (OASIS). Depression and anxiety. 2006;23(4):245-9.
18. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association. 2013 Jan 1;20(1):144-51.
19. Payrovnaziri SN, Xing A, Salman S, Liu X, Bian J, He Z. Assessing the impact of imputation on the interpretations of prediction models: A case study on mortality prediction for patients with acute myocardial infarction. AMIA Summits on Translational Science Proceedings. 2021;2021:465.
20. Wu J, Roy J, Stewart WF. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Medical care. 2010 Jun 1:S106-13.
21. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002 Jun 1;16:321-57.
22. Hernandez-Boussard T, Bozkurt S, Ioannidis JP, Shah NH. MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. Journal of the American Medical Informatics Association. 2020 Dec;27(12):2011-5.
23. Lee S, Lee CW. Application of decision-tree model to groundwater productivity-potential mapping. Sustainability. 2015 Sep 30;7(10):13416-32.
24. Svoray T, Michailov E, Cohen A, Rokah L, Sturm A. Predicting gully initiation: comparing data mining techniques, analytical hierarchy processes and the topographic threshold. Earth Surface Processes and Landforms. 2012 May;37(6):607-19.
25. Akar Ö, Güngör O. Classification of multispectral images using Random Forest algorithm. Journal of Geodesy and Geoinformation. 2012 Nov;1(2):105-12.
26. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural networks. 2005 Jul 1;18(5-6):602-10.
27. Azuaje F, Dubitzky W, Black N, Adamson K. Improving clinical decision support through case-based data fusion. IEEE Transactions on biomedical engineering. 1999 Oct;46(10):1181-5.
28. Joshi G, Walambe R, Kotecha K. A review on explainability in multimodal deep neural nets. IEEE Access. 2021 Mar 31;9:59800-21.
29. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014 Dec 22.
30. Venugopalan J, Tong L, Hassanzadeh HR, Wang MD. Multimodal deep learning models for early detection of Alzheimer's disease stage. Scientific reports. 2021 Feb 5;11(1):1-3.
31. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. InProceedings of the IEEE international conference on computer vision 2017 (pp. 2980-2988).
32. Dietterich TG. Ensemble methods in machine learning. InInternational workshop on multiple classifier systems 2000 Jun 21 (pp. 1-15). Springer, Berlin, Heidelberg.
33. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, Liu X, He Z. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. Journal of the American Medical Informatics Association. 2020 Jul;27(7):1173-85.
34. Jeng JS, Huang SJ, Tang SC, Yip PK. Predictors of survival and functional outcome in acute stroke patients admitted to the stroke intensive care unit. Journal of the neurological sciences. 2008 Jul 15;270(1-2):60-6.
35. Ho WM, Lin JR, Wang HH, Liou CW, Chang KC, Lee JD, Peng TY, Yang JT, Chang YJ, Chang CH, Lee TH. Prediction of in-hospital stroke mortality in critical care unit. Springerplus. 2016 Dec;5(1):1-9.