College Student Retention Risk Analysis from Educational Database Using Multi-Task Multi-Modal Neural Fusion

Mohammad Arif Ul Alam

University of Massachusetts Lowell, MA, USA
mohammadariful_alam@uml.edu

Abstract

We develop a Multimodal Spatiotemporal Neural Fusion network for Multi-Task Learning (*MSNF-MTCL*) to predict 5 important students' retention risks: future dropout, next semester dropout, type of dropout, duration of dropout and cause of dropout. First, we develop a general purpose multimodal neural fusion network model *MSNF* for learning students' academic information representation by fusing spatial and temporal unstructured advising notes with spatiotemporal structured data. *MSNF* combines a Bidirectional Encoder Representations from Transformers (BERT)-based document embedding framework to represent each advising note, Long-Short Term Memory (LSTM) network to model temporal advising note embeddings, LSTM network to model students' temporal performance variables and students' static demographics. The final fused representation from *MSNF* has been utilized on a Multi-Task Cascade Learning (*MTCL*) model towards building *MSNF-MTCL* for predicting 5 student retention risks. We evaluate *MSNF-MTCL* on a large educational database consists of 36,445 college students over 18 years period of time that provides promising performances comparing with the nearest state-of-the-art models. Additionally, we test the fairness of such model given the existence of biases.

Introduction

The U.S. National Center for Education Statistics (NCES) reports that in United States, the average retention rate for higher education institutions is 71% (McFarland 2017). While, 57% of college admitted students do not complete four-year colleges within six years, 33% of them drop out from college without any degree (McFarland 2017). For some students, dropping out is the culmination of years of academic hurdles, missteps, and wrong turns. For others, the decision to drop out is a response to conflicting life pressures, the need to help support their family financially or the demands of caring for siblings or their own child. Dropping out is sometimes about students being bored and seeing no connection between academic life and "real" life. It's about young people feeling disconnected from their peers and from teachers and other adults at school (CATERALL 1998). Although the reasons for dropping out vary, the consequences of the decision are remarkably similar. Low retention rates not only impact the financial well-being of individuals but the economy as a whole, college dropouts are more likely to head down a path that leads to lower-paying jobs, poorer health, and the possible continuation of a cycle of poverty that creates immense challenges for families, neighborhoods, and communities (McFarland 2017). Low retention rates also adversely affect the reputation of the educational institution and could lead to potential loss of funding and inability to compete for quality students (CRONINGER and LEE 2001). Thus, improving student retention is of paramount importance at institutions of higher education.

Many researchers have proposed to model factors impacting student dropout from large scale educational database using statistical and machine learning models. Most researchers have focused on using static or temporal structured data, such as GPA, SAT scores etc., that are readily available in institutional databases (Prenkaj, Stilo, and Madeddu 2020a). Some of the researchers proposed to use unstructured text analysis such as advising notes, forum post, social media status, online chats and email mining using natural language processing techniques to predict student dropout (Jayaraman 2020; Tinto 1993). However, none of the researchers proposed to combine structured and unstructured data in spatiotemporal fashion that can provide significant promise in this domain of research. We propose *MSNF-MTCL* with the following **key contributions**:

- We develop a novel multimodal spatiotemporal neural fusion model *MSNF* for educational database to fuse temporal student advising notes extracted BERT embedding, temporal student performance variables and static student demographic information via temporal document encoder, temporal performance encoder and static demographic encoder respectively.
- We develop a cascaded information network-based Multi-Task Cascade Learning (MTCL) layer on the top of the fusion layer to build our core *MSNF-MTCL* model by placing lower-level tasks at earlier layers so that the features learned for these tasks may be used by higher-level tasks for 5-tasks MTCL problem.
- We evaluate *MSNF-MTCL* on a large scale collected data from a University from a third world country via comparing the performance with nearest state-of-the-art solutions.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

- Additionally, we tested the existence of biases and applied bias mitigation technique to confirm fairness of *MSNF-MTCL*.

Related Works

Traditionally, education researchers run surveys to find the facts impacting dropped out students dropout that include academic difficulty, adjustment problems, lack of clear academic goals, lack of commitment, inability to integrate with the college community, uncertainty, incongruence, isolation as factors involved in student dropout (Tinto 1993). The surveys result some key factors such as past and current academic success, high school GPA, SAT scores (Porter 2008), major and number of credit hours taken during the first semester (Cabrera 1993). effect of financial aid (Herzog 2005). Machine learning techniques on educational database has been relatively new (Alfredo Perez 2018; Iamon and Boongoen 2017; Prenkaj, Stilo, and Madeddu 2020b; Coussement et al. 2020; Pellagatti et al. 2021). Perez et. al. proposed logistic regression and decision tree based dropout prediction from static students' data (Alfredo Perez 2018). (Iam-on and Boongoen 2017) proposed a link-based cluster ensemble for predicting student dropout from mixed-type (categorical and continuous) educational dataset. (Prenkaj, Stilo, and Madeddu 2020b) presented benchmark student dropout definition and dropout prediction paradigm by developing machine and deep learning techniques and their related privacy concerns from static and temporal structured data. (Coussement et al. 2020) proposed logit leaf model (LLM) on students classroom characteristics, cognitive and behavioral engagement variables and other static variables available from online students' enrollment database. (Pellagatti et al. 2021) proposed a Generalized mixed-effects random forest model to analyze hierarchical data to predict engineering students' dropout from static data from large scale educational dataset. On the other hand, student dropout prediction from advising notes has been explored only once (Jayaraman 2020) that proposed a sentiment analysis technique to mine advising notes towards predicting students' dropout. Additionally, this paper proposed an explanation i.e. weighted ranking of contributing sentiments towards predicting students' dropout. (Yu, Lee, and Kizilcec 2021) proposed a fair student dropout prediction system from educational database. (Prenkaj, Stilo, and Madeddu 2020a) analyzed the challenges of student dropout from static database that involves definition, machine learning techniques to be used, evaluation measures and privacy concerns.

Combining structured and unstructured data has been popular in image processing content learning and electronic health record analytics for decades. (Wan et al. 2014) proposed deep spatial CNN model to extract features from image-text pairs. (Tam et al. 2021) presents LSTM-CNN fusion to combine clinical image and electronic health records together for predicting clinical events derived cohorts. (Wu 2021) presents utilized unstructured-structured text fusion model for predicting cognitive engagement. Similar approach has been conducted in many domains such as mortality prediction (Baxter et al. 2020), structured visualization from unstructured texts (Li et al. 2021), financial transaction prediction (Au, Ait-Azzi, and Kang 2021) and so on.

Multi-task learning (MTL) has been investigated mostly by computer vision researchers that are categorized in many terms such as shared trunk, cross-talk, prediction distilation, task routing. In NLP, the MTL falls under many categories. Traditional feed-forward neural networks (non-attention based) focused on developing structural resemblance of shared global feature extractor followed by task-specific output branches where features are word representations (Collobert et al. 2011). Recurrent neural network models in MTL mostly focused on novel recurrent neural architectures adopted in multi-task fashion with multi-variant parameter sharing schemes i.e., one-to-one, one-to-many and many-to-many or task specific LSTMs (Liu et al. 2015; Dong et al. 2015). Cascaded information techniques mostly focused on lower-level tasks at earlier layers so that the features learned for these tasks may be used by higher-level tasks (Sanh, Wolf, and Ruder 2019). Adversarial feature separation techniques introduce an adversarial learning framework for MTL in order to distill learned features into task-specific and task-agnostic subspaces. Their architecture is comprised of a single shared LSTM layer and one task-specific LSTM layer per task(Ruder et al. 2019). BERT in MTL mostly focused on adding shared BERT embedding layers on the traditional, LSTM or cascaded information technique (Liu et al. 2020).

To our best knowledge, *(MSNF-MTCL)* is the first of its kind, that develops a Multimodal Spatiotemporal Neural Fusion for MTL model combining structured, unstructured, spatio-temporal contexts on educational data. More elaborately, we design multimodal neural network model to fuse static students' structured demographic information, temporal students' structured performance information and temporal students' unstructured advising notes and develop a novel classification model towards predicting student dropout, next semester dropout and dropout cause identification.

Data Description

We obtained an educational database from a private university located in a developing world country consists of 36,445 undergraduate students where female (10,237) and male (26,208) students' ratio (28% by 72%) is similar to national literacy statistics of the country. Among the students, 14% are dropped out (female-male dropout are 11% and 15%) in any point of their study. While any dropout incident happened, dropped out students were contacted by university counselling office via phone to analyze the incident which has been categorized into two classes (1) temporary dropout, (2) permanent dropout. Here, the causes of permanent dropout has been sub-categorized into 10 classes (financial, family, marriage, sickness and so on) and temporary dropout has been sub-categorized into 14 classes (financial, internship, sickness, accident, marriage, COVID-19 related, family member death, struggling with grades and so on). Both of temporary and permanent dropout causes have 9 overlaps and in total 15 unique causes have been structured to represent any kind of dropout causes. It should be noted that location transfer and university transfer reasons

Gender	Count	Dropout	Temporary	Permanent
Female	5,498 (24%)	1,103 (11%)	717 (65%)	386 (35%)
Male	17,897 (76%)	3,857 (15%)	2,931 (76%)	926 (15%)
Total	23,395	4,960 (14%)	3,648 (74%)	1,312 (26%)

Table 1: Description of the obtained educational database

were not considered as dropout in the inclusion criteria, and these information has been removed from every statistics. While getting admitted, students were provided few demographic data related to students personal profile, prior education details, family information and financial information. Since the admission, university administration has been recording students' temporal performances in each courses taken along with few administrative structured information such as payment due, blocked to register for next semester (due to any critical incidents, past significant payment dues), scholarship awarded etc. Each semester, students were required to visit to his/her academic advisor to discuss various topics related to academia which is more likely to be the first month of the semester. Sometimes, students were blocked from registering to next semester without consulting academic advisor due to many reasons, such as, poor grades, excessive missing of attendance, payment dues. However, students also could schedule meeting with their academic advisor anytime of the semester to discuss various topics (from personal to academic). It should be noted that, only primary cause of dropout has been noted during the counselling session. Table 1 and Table 2 present the details of the statistics of the dataset and features information derived/extracted from the database respectively.

Variables	Features
Static and structured Demographic	birth date, age, gender, religion, starting major, transferred credits, blood group, birth place, permanent address, local address, Secondary School grade, higher school grade, marital status, source of finance, part-full time, local guardian, parents financial income
Temporal and structured Performance	new credits taken, credits retaken, passing credits, failed credits, overall attendance, average semester starting GPA, average semester GPA, average semester ending GPA, number of exams unattended since admission, number of exams unattended in this semester, number of counselling scheduled, amount of payment due in this semester, number of payment dues since admission, study duration, blocked from registering in next semester, number of block since admission, scholarship amount, accommodation status (on/off campus), total scholarship till date, average scholarship per semester
Temporal Advising Notes	*structured:* reason of counselling visits, counselling conduct date, counselling result (no result or cause of dropout) *unstructured:* counselling note
Dropout Causes	*financial, *family, *marriage, *physically ill, *death of family member, *personal, †death, *accident, *struggling with grades, *COVID-19 family death, COVID-19 financial, COVID-19 online class attending hardship, internship, traveling, mentally ill

Table 2: Description of the features provided/generated from the educational database

Multi-Task Multi-Modal Neural Fusion Model for Predicting Student Retention Risks

In this section, we describe the problem formulation, multimodal spatiotemporal neural fusion and multi-task neural cascade networks to solve student retention risks prediction. The overall framework has been shown in Fig 1. The lower module "Multi-Modal Fusion" generates, L, a spatiotemporal fused layer that has been shared across all tasks, while the upper module "Multitask Neural Cascades" represent task-specific outputs, L, in our case $L \in \{L_1, L_2, L_3, L_4, L_5\}$.

Multimodal Spatiotemporal Neural Fusion

This module consists of advising note representation via BERT-based document embedding, sequential encoder network on temporal advising note documents from BERT embedding, development of temporal structured performance information encoder, development of static information encoder and a fusion layer that has been shared by each task of Multi-Task Cascade network. The input can be represented as $X \in \{P, D, N\}$ where P, D, N represent temporal structured performance data, static students demographic data and temporal students' advising/counselling notes. The output of this layer is fused representation of spatiotemporal inputs of X which can be represented as $Z \in \{Z_{temporal}, Z_{static}, Z_{note}\}$.

Static Information Encoder (Z_{static}) The static student description (Table 2) data $D \in \{D_1, D_2..D_n\}$ has been converted into one-hot vectors through static student description encoder to generate output Z_{static}. This encoder consists of a series of convolution (CNN) layers, where each CNN layer further followed by batch normalization, max pooling, and dropout layer. The first 1D CNN layer takes the one-hot encoded static feature and structured demographic data (size: 120) as input and performs the filter operation with 8 filters of size 11. The outputs of the first CNN layer are passed to the second CNN layer (16 filters with a size of 5). Next, the outputs of the second CNN layer are passed to the third CNN layer (32 filters with a size of 3). Finally, the summary of all the spatial features of a static input feature is passed to the flatten layer to produce a 1D feature vector of size 50.

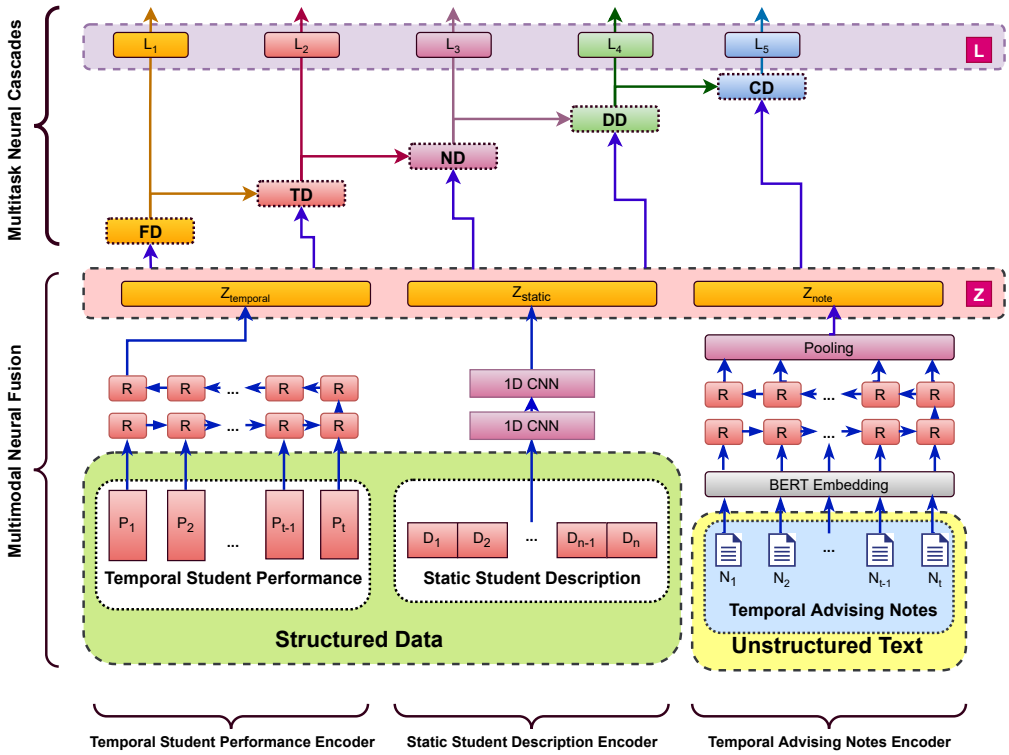

Figure 1: Overall Architecture of Multimodal Spatiotemporal Neural Fusion (MSNF) Network model for predicting student dropout risks i.e. dropout, next semester dropout and cause of dropout

Temporal Student Performance Encoder ($Z_{temporal}$) To capture the longer dynamics in the temporal dimension of the temporal student performance data, $P \in \{P_1, P_2..P_t\}$ where t represents the time, we have used two consecutive LSTM layers: The first LSTM layer consists of 75 neurons, and the second one with 55 neurons. Each LSTM layers are followed by a dropout and batch normalization layer. Next, a dense layer of 50 neurons, followed by a dropout and batch normalization layer, is connected to another dense layer with 40 neurons. Finally, informative features of the input P have been extracted to generate final encoded layer $Z_{temporal}$.

Sequential Advising Note Encoder (Z_{note}) The input $N \in \{N_1, N_2..N_t\}$ here t represents time, which is a document sequence. At first we perform BERT pre-trained embedding fine-tuning as proposed in (Devlin et al. 2019). First, we consider, each of the document consists of a sequence of sentences and each sentence has been considered as a sequence of words. We represent each of the sequence of word separated by token [CLS] while each sequence of sentence has been separated by [SEP] token as described in (Devlin et al. 2019) proposed method. Then we map the final tokenized document into a sequence of input embedding vectors, one for each token, constructed by summing the corresponding word, segment, and positional embeddings, thus it is called input representation vector. Now, we use multi-layered bidirectional Transformer encoder (BERT) (Devlin et al. 2019) pre-trained embedding to map input representation vectors into a sequence of contextual embedding vectors. Then, the sequence of contextual embedding vectors are passed through a Bidirectional LSTM (BiLSTM) (Zhang et al. 2020). The BiLSTM layer concatenates the outputs from 2 hidden layers of opposite direction to the same output and can capture long term dependencies in sequential text data. The maxpooling layer takes the hidden states of the BiLSTM layer as input and outputs the final text representation Z_{note} (Zhang et al. 2020).

Student Spatiotemporal Information Representation (Z) The final students' spatiotemporal information representation Z is obtained by concatenating the representations of sequential advising note, temporal student performance, along with static student demographic information. The representation of each student is $z_p \in Z = [Z_{temporal}, Z_{static}, Z_{note}]$ the size of this vector is $d_{temporal}, d_{static}, d_{note}$.

Multi-Task Neural Cascade Networks

We leverage the final task L as hierarchical composition of five tasks ($L \in \{L_1, L_2, L_3, L_4, L_5\}$) for future dropout, type of dropout, next semester dropout, duration of dropout and cause of dropout tasks respectively, to train our student retention risk predictor by developing a Multimodal Spatiotemporal Neural Fusion network for MTL (*MSNF-MTCL*). We formulate two types of losses:

- Categorical cross-entropy loss for classification task

$$L_i^{cat} = -(y_i^{cat} log(p_i) + (1 - y_i^{cat})(1 - log(p_i)) \quad (1)$$

where p_i denotes probability of the classification task and $y_i^{fd} \in \{y_1, .., y_n\}$ denotes the ground-truth labels.

- Euclidean loss for regression task

$$L_i^{reg} = ||\hat{y}_i^{reg} - y_i^{reg}||_2^2 \quad (2)$$

where $\hat{y}_i^{reg}$ is the continuous estimated regression task values and y_i^{reg} is the ground truth.

We define each of task as of our multi-task model along with the final multi-source learning scheme as follows:

Future Dropout (FD) This is a binary task involves predicting students' dropout in future (true/false) which is irrespective of the semester or duration. The learning objective is formulated as a two-class classification problem. For each sample, we use the cross-entropy loss L_i^1 similar to Eqn. 1 where where p_i is probability of dropout in future and $y_i^1 \in \{0, 1\}$ denotes ground truth label.

Type of Dropout (TD) This binary task aiming to further categorize dropout into temporary or permanent. Similar to Eqn 1, we can formulate L_i^2 where where p_i is probability of type of dropout (temporary dropout, permanent dropout and $y_i^2 \in \{0, 1\}$ denotes ground truth label.

Next Semester Dropout (ND) This binary task aims to predict whether predicted dropped out student will be dropped out in next semester or not. We use the cross-entropy loss L_i^3 similar to Eqn. 1 where where p_i is probability of next semester dropout and $y_i^3 \in \{0, 1\}$ denotes ground truth label.

Duration of Dropout (DD) This regression task aims to predict how many semesters students survive if the dropout has been predicted. We use the Euclidean loss L_i^4 similar to Eqn. 2 where where $\hat{y}_i^5$ is the continuous estimated duration of dropout in terms of semester and y_i^5 is the ground truth.

Cause of Dropout (CD) This task aims to predict the causes of dropout, i.e. one of the 15 causes as stated in Table 2. We use the cross-entropy loss L_i^5 similar to Eqn. 1 where where p_i is probability of each cause of dropout and $y_i^5 \in \{0, 1, .., 14\}$ denotes ground truth label.

Multi-Conditional Training We employ five different tasks on our encoded students' information space Z, there are different types of labels in each training sample. While training on the samples, we follow the hierarchy of $L_1(FD) \rightarrow L_2(TD) \rightarrow L_3(ND) \rightarrow L_4(DD) \rightarrow L_5(CD)$ and develop an overall learning target as follows

$$L(\Theta) = L_1 + L_2 + L_3 + L_4 + L_5 \quad (3)$$

While computing L, we abide the following strategies: if $y_i^1 = 0$ (no dropout), then we set, $L_2 = L_3 = L_4 = L_5 = 0$, if $y_i^1 = 1$ (no dropout) and $y_i^2 = 0$ (permanent dropout), then we set, $L_3 = L_4 = L_5 = 0$, , if $y_i^1 = 1$ (no dropout), $y_i^2 = 0$ (permanent dropout), and $y_i^3 = 0$ (next semester dropout = true), then we set, $L_4 = L_5 = 0$. We compute L considering altogether as per Eqn. 3 for all other cases.

Experiments

Baseline Models

Since, multi-task multi-modal neural fusion on educational dataset is a novel problem for student retention risks estimation, we could not find state-of-the-art solutions that match with our problem as a baseline. In this regard, we implement few nearest problems along with their solutions and formulate similar problem using our proposed *MSNF-MTCL* framework. Apart from that, to establish the importance of different modules of our framework, we develop different versions of *MSNF-MTCL* consist of different combinations of proposed modules. The baselines and different versions of *MSNF-MTCL* framework have been described below:

- **B1 (Jayaraman Model) (Jayaraman 2020):** This framework utilized only advising note and proposed a lexicon-based sentiment analysis technique to extract features and applied SVM machine learning techniques on the features to predict student dropout. The framework utilized Bing Lexicon (Liu 2010) model for feature extraction that consists of 6,800 words, 2,000 positive and 4,800 negative sentiments.
- **B2 (Pellagatti Model) (Pellagatti et al. 2021):** This framework considered students' static and students' temporal structured data towards building a generalized mixed-effects random forest (GMERF).
- **B3 (Single Task Fusion and Replacing BERT with Doc2Vec) (Zhang et al. 2020):** This framework is the closest one to our solution that has been developed to predict mortality of patients from electronic health records (EHR). It followed a spatiotemporal neural fusion of patient notes, patients' static demographic data and patient's temporal hospital information altogether into a fused layer that has been utilized to solve single task, predicting patients' mortality. Instead of using lexicon tokenization and BERT model for encoding patient notes, this framework utilized Doc2Vec embedding (Le and Mikolov 2014).
- **V1 (MSNF-MTCL with Structured Data Only):** This is a version of our proposed core *MSNF-MTCL* model where we completely removed Temporal Advising Notes input and considered only Structured data i.e. Temporal Student Performance and Static Student Description inputs along with their encoders.
- **V2 (MSNF-MTCL with Unstructured Advising Notes Only):** This is a version of our proposed core *MSNF-MTCL* model where we included only Temporal Advising Notes input and its corresponding encoder.
- **V3 (MSNF-MTCL):** This is a complete *MSNF-MTCL* model including all modules and inputs.

Results

We considered accuracy $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ $where\ TP, TN, FP\ FN\ denote\ true-positive, true\ -\ negative, false\ -\ positive, false\ -\ negative$ and Standard Deviation $\pm\%$ as evaluation metric for classification tasks. We considered root mean squared

Data #Classes	B1	B2	B3	V1	V2	V3 (Ours)
FD (2)	72.45 ± 9.3	73.51 ± 8.6	75.47 ± 6.8	80.76 ± 3.8	82.84 ± 4.2	**98.78±0.01**
TD (2)	61.65 ± 8.8	66.56 ± 9.1	70.77 ± 8.4	76.42 ± 4.3	79.54 ± 5.3	**89.73±0.01**
ND (2)	60.65 ± 10.4	65.63 ± 9.2	68.84 ± 8.3	78.73 ± 4.2	80.25 ± 4.1	**93.25±0.01**
DD	5.35 ± 0.87	3.65 ± 0.66	2.3 ± 0.56	1.1 ± 0.18	0.85 ± 0.05	**0.045±0.002**
CD (15)	59.83 ± 8.4	59.42 ± 9.3	60.27 ± 11.53	68.54 ± 5.3	70.54 ± 3.5	**85.53±0.02**

Table 3: Comparison of *MSNF-MTCL* performance on our dataset with different baseline models

	SPD	EOD	AOD	DI	FD	ND	TD	DD	CD
Fairness target	-0.1 to 0.1	-0.1 to 0.1	-0.1 to 0.1	0.8 to 1.2	98.78	89.73	93.25	0.045	85.53
Initial	0.25	-0.18	-0.19	0.53					
RW	**0.05**	**-0.03**	-0.15	**0.95**	91.85	86.73	90.55	0.223	81.47
AB	0.09	**-0.07**	-0.11	**1.0**	90.34	87.45	89.65	0.23	81.34
ROBC	**0.06**	-0.11	**0.08**	**0.91**	91.24	86.43	89.38	0.09	80.44
EOPP	0.18	-0.15	**-0.07**	**0.88**	90.75	87.47	85.76	0.23	80.43
DIR	**0.06**	**-0.09**	-0.11	**0.11**	88.36	85.83	86.99	0.24	83.05
LFR	0.20	-0.10	**0.01**	**1.0**	90.77	83.84	88.87	0.145	82.75
CEOP	**0.05**	**-0.05**	-0.11	**0.89**	89.76	85.4	90.93	**0.049**	80.34
PR	**0.06**	**-0.09**	**-.04**	**0.91**	**93.53**	**88.83**	**92.54**	0.055	**83.46**

Table 4: Bias detection and mitigation experiment results. Here, column represents bias detection metrics: Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD), Average Odds Difference (AOD) and Disparate Impact (DI); while rows represent bias mitigation techniques: Reweighing (RW), Adversarial Debiasing (AB), Reject Option Based Classification (ROBC), Equalized odds post processing (EOPP), Disparate impact remover (DIR), Learning fair representation (LFR), Calibrated equalized odds postprocessing (CEOP) and Prejudice remover (PR) for each of the task: future dropout (FD), next semester dropout (ND), type of dropout (TD), duration of dropout (DD) and cause of dropout (CD)

deviation (RMSD) as evaluation metric for regression tasks. We implemented baseline algorithms and our framework using python-based Keras library. We train the model using a learning rate of 0.001 for 16k iterations, and 0.0001 for the next 5k until the training converges. We train the model in 4 GPUs, each GPU holding 1 mini-batch (so the effective mini-batch size is x4).

While developing baseline algorithms, we designed 5 single task models for 5 retention risks. We considered 75% of students' data as training and rest of 25% of students' data as testing data during training and similar experiment has been conducted 10 times on 10-fold cross experiment to generate the results. We also utilized Synthetic Minority Oversampling Technique (SMOTE) to correct the imbalance (Finlay, Pears, and Connor 2014). SMOTE is a popular and robust technique that uses a combination of oversampling the minority class and undersampling the majority class which results in better classifier performance than just oversampling or undersampling. Table 3 shows detail results of our experiment and comparisons.

In Table 3, we clearly can see that our proposed method (V3-Ours) perform better than any other baseline frameworks (B1, B2 or B3) for all student retention risk classification/estimation. If we take closer look, we can see that, utilizing only advising notes (V1) and only structured data (V2) versions of our framework not only outperform their related baselines (only advising note B1 and only structured data B2), the outperform state-of-the-art single task spatiotem-

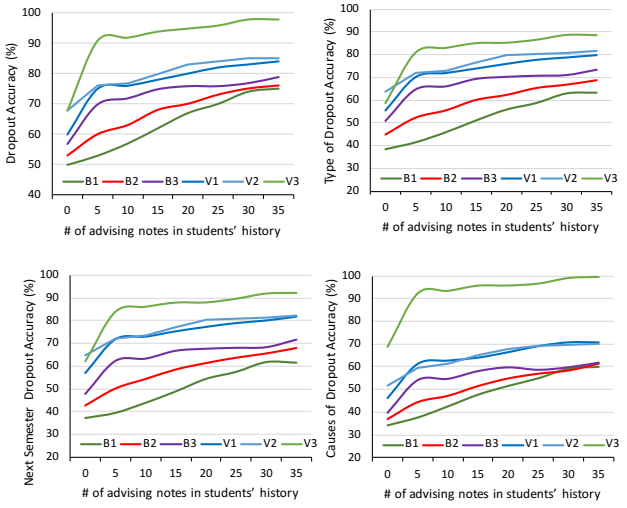

Figure 2: Accuracy changes of five different retention risk prediction tasks using our framework over number of available advising notes

poral fusion model using Doc2Vec embedding framework which has been successfully applied on EHR data before.

Bias Detection and Mitigation

Table 1 shows that the data is biased in terms of gender (female-male ratio is 28% by 72%) which has potential threat to AI fairness in our model. We utilize IBM AI Fairness 360 (AIF360) tool to detect and mitigate biases for dropout prediction in terms of gender considering "Male" as privileged group (Bellamy et al. 2019). Table 4 shows AIF360 implemented 4 bias detection metrics, their corresponding fairness target metric ranges and 8 bias mitigation techniques generated bias detection metrics. The central notions in this method: (1) all bias mitigation techniques are not appropriate for every dataset; (2) to select right mitigation technique, the bias detection metrics should be fair under maximum metrics; (3) accuracy drop due to bias mitigation should be minimum. Table 4 shows the final result of our bias detection and mitigation test for student dropout (only the first task of our multi-task model) where we can see that "Prejudice remover" technique provides maximum fairness (fair in 4 bias detection metrics) and least accuracy drop (accuracy drop of 3.33%).

Discussion

Fig. 2 illustrates the changes of accuracies over the number of availability of each student's advising note while predicting their retention risks (five different tasks) which clearly shows that different versions of our method (V1, V2 and V3) outperform baseline methods significantly in any number of advising notes' availability. Also, it can be clearly stated that, the prediction accuracy of each task increases as the number of available advising notes increases for each student in the testing data. Fig 3 illustrates the prediction accuracies of individual dropout cause (15 dropout causes) using our proposed model, where we can see that (we removed cause "Own Death with index 7" due to ethical reason), predicting dropout due to financial condition, family reason, marriage related, struggling with grades, COVID-19 related financial, COVID-19 related struggling in attending online classes and mentally ill, are extremely accurate (95%+). However, it has been extremely difficult to predict physical illness, death of family member and personal problem related college dropout from the educational data.

Limitations and Future Work

We utilized a large scale educational data of 18 years from only one university which may create distribution biases. To address biases, we additionally tested our framework for bias mitigation. Moreover, our reproduction of baseline models and evaluation on our dataset provide ample proof that our model outperforms baseline frameworks. In our framework, lower level cascaded task depends on the performance of upper level tasks' classification performances that we did not align with state-of-the-art models' implementations. The causes of dropout have been labeled in rolling basis, i.e., when a faculty advisor thought that current advisee needs to be assigned to a new cause, he reports to the system for an additional cause insertion. The administration officer review that cause and accept the inclusion request if that is absolutely valid. Our dataset consists of pre- and post- COVID-19 pandemic data. However, due to extremely poor number of data during post-COVID-19 era, we could not develop a new model to identify COVID-19 impacts on student dropout. In the current system, a faculty advisor can only assign a single cause for a single advising note, that made us difficult to predict multiple causes of a dropout incident which is common in real life case. In future, we aim to apply causal inference and information retrieval technique for facts finding to describe COVID-19 impacts and multiple causes extraction on student dropout more evidently. We also utilized pre-trained BERT embedding model that has been trained on wikipedia data. In future, we plan to develop a new embedding, "Educational BERT (EBERT)" trained on only educational advising notes to enhance efficiency of any student retention risk prediction.

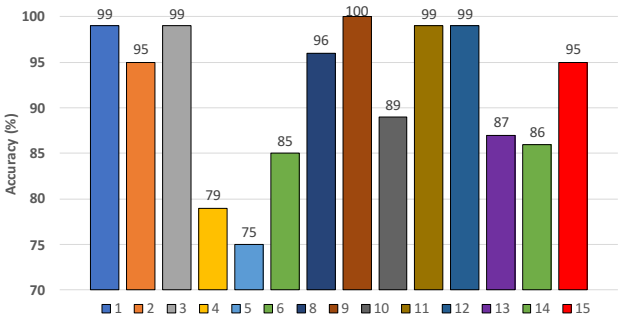

Figure 3: Causes of Dropout prediction results using our overall framework. The causes of dropout have been indexed with: 1. financial, 2. family, 3. marriage, 4. physically ill, 5. death of family member, 6. personal, 7. own death, 8. accident, 9. struggling with grades, 10. COVID-19 family death, 11. COVID-19 financial, 12. COVID-19 online class attending hardship, 13. internship, 14. traveling, 15. mentally ill. We removed 7.own death due to ethical reason.

Conclusion

Structured-unstructured data fusion in spatiotemporal domain across the educational institute has not been properly exploited by researchers due to the unavailability of such data and challenges of combining multi-modal educational signals. Our breakthrough approach that provides highest ever student dropout accuracy potentially can be adopted by educational policy makers and university management stakeholders in many other domains. Our novel problem formulation, a multi-task student retention risks estimation on 5 different student retention risk tasks, and solution, an efficient multi-task multi-modal spatiotemporal neural network model will open the door to many unsolved problems in educational data mining research. Moreover, the framework can be adapted in any databases in the world including employee, email, electronic health record or google search databases, and, can be utilized to solve extremely complex problems.

References

Alfredo Perez, M. C. G. V., Elizabeth E. Grandon. 2018. Comparative Analysis of Prediction Techniques to Determine Student Dropout: Logistic Regression vs Decision Trees. *SCCC*.

Au, W.; Ait-Azzi, A.; and Kang, J. 2021. FinSBD-2021: The 3rd Shared Task on Structure Boundary Detection in Unstructured Text in the Financial Domain. In *WWW (Companion Volume)*, 276–279. ACM / IW3C2.

Baxter, S. L.; Klie, A. R.; Saseendrakumar, B. R.; Ye, G. Y.; Hogarth, M. A.; and Nemati, S. 2020. Predicting Mortality in Critical Care Patients with Fungemia Using Structured and Unstructured Data. In *EMBC*, 5459–5463. IEEE.

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J. T.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.*, 63(4/5): 4:1–4:15.

Cabrera, N. A. . C. M. B., A. F. 1993. College persistence: Structural equations modeling test of an integrated model of student retention. In *The Journal of Higher Education*.

CATERALL, J. S. 1998. Risk and Resilience in Student Transition to High Schools. *American Journal of Education*.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. P. 2011. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.*, 12: 2493–2537.

Coussement, K.; Phan, M.; Caigny, A. D.; Benoit, D. F.; and Raes, A. 2020. Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decis. Support Syst.*, 135: 113325.

CRONINGER, R.; and LEE, V. E. 2001. Social Capital and Dropping out of High School: Benefits to At-Risk Students of Teachers' Support and Guidance. *Teachers College Record*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

Dong, D.; Wu, H.; He, W.; Yu, D.; and Wang, H. 2015. Multi-Task Learning for Multiple Language Translation. In *ACL (1)*, 1723–1732. The Association for Computer Linguistics.

Finlay, J.; Pears, R.; and Connor, A. M. 2014. Synthetic Minority Over-sampling TEchnique (SMOTE) for Predicting Software Build Outcomes. In *SEKE*, 546–551. Knowledge Systems Institute Graduate School.

Herzog, S. 2005. Measuring Determinants of Student Return vs. Dropout/Stopout vs. Transfer: A First-to-Second Year Analysis of New Freshmen. *Research in Higher Education*, 46(8): 883–928.

Iam-on, N.; and Boongoen, T. 2017. Improved student dropout prediction in Thai University using ensemble of mixed-type data clusterings. *Int. J. Mach. Learn. Cybern.*, 8(2): 497–510.

Jayaraman, J. D. 2020. Predicting Student Dropout by Mining Advisor Notes. In *EDM*. International Educational Data Mining Society.

Le, Q. V.; and Mikolov, T. 2014. Distributed Representations of Sentences and Documents. In *ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, 1188–1196. JMLR.org.

Li, T.; Fang, L.; Lou, J.; Li, Z.; and Zhang, D. 2021. AnaSearch: Extract, Retrieve and Visualize Structured Results from Unstructured Text for Analytical Queries. In *WSDM*, 906–909. ACM.

Liu, B. 2010. Sentiment Analysis and Subjectivity. In *Handbook of Natural Language Processing*, 627–666. Chapman and Hall/CRC.

Liu, X.; Wang, Y.; Ji, J.; Cheng, H.; Zhu, X.; Awa, E.; He, P.; Chen, W.; Poon, H.; Cao, G.; and Gao, J. 2020. The Microsoft Toolkit of Multi-Task Deep Neural Networks for Natural Language Understanding. In *ACL (demo)*, 118–126. Association for Computational Linguistics.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*, 3730–3738. IEEE Computer Society.

McFarland, H. B. d. B. C. S. T. W. X., J. 2017. The Condition of Education. *National Center for Education Statistics*.

Pellagatti, M.; Masci, C.; Ieva, F.; and Paganoni, A. M. 2021. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. *Stat. Anal. Data Min.*, 14(3): 241–257.

Porter, K. B. 2008. Current trends in student retention: A literature review. *Teaching and Learning in Nursing*, 3(1): 3–5.

Prenkaj, B.; Stilo, G.; and Madeddu, L. 2020a. Challenges and Solutions to the Student Dropout Prediction Problem in Online Courses. In *CIKM*, 3513–3514. ACM.

Prenkaj, B.; Stilo, G.; and Madeddu, L. 2020b. Challenges and Solutions to the Student Dropout Prediction Problem in Online Courses. In *CIKM*, 3513–3514. ACM.

Ruder, S.; Bingel, J.; Augenstein, I.; and Søgaard, A. 2019. Latent Multi-Task Architecture Learning. In *AAAI*, 4822–4829. AAAI Press.

Sanh, V.; Wolf, T.; and Ruder, S. 2019. A Hierarchical Multi-Task Approach for Learning Embeddings from Semantic Tasks. In *AAAI*, 6949–6956. AAAI Press.

Tam, C. S.; Gullick, J.; Saavedra, A.; Vernon, S. T.; Figtree, G. A.; Chow, C. K.; Cretikos, M.; Morris, R. W.; William, M.; Morris, J.; and Brieger, D. 2021. Combining structured and unstructured data in EMRs to create clinically-defined EMR-derived cohorts. *BMC Medical Informatics Decis. Mak.*, 21(1): 91.

Tinto, V. 1993. Building community. In *Liberal Education*. Liberal Education.

Wan, J.; Wang, D.; Hoi, S. C. H.; Wu, P.; Zhu, J.; Zhang, Y.; and Li, J. 2014. Deep Learning for Content-Based Image Retrieval: A Comprehensive Study. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM

'14, 157–166. New York, NY, USA: Association for Computing Machinery. ISBN 9781450330633.

Wu, J. 2021. Learning analytics on structured and unstructured heterogeneous data sources: Perspectives from procrastination, help-seeking, and machine-learning defined cognitive engagement. *Comput. Educ.*, 163: 104066.

Yu, R.; Lee, H.; and Kizilcec, R. F. 2021. Should College Dropout Prediction Models Include Protected Attributes? *CoRR*, abs/2103.15237.

Zhang, D.; Yin, C.; Zeng, J.; Yuan, X.; and Zhang, P. 2020. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Medical Informatics Decis. Mak.*, 20(1): 280.