Article

Not peer-reviewed version

Towards Predicting Length of Stay and Identification of Cohort Risk Factors Using Self-Attention Based Transformers and Association Mining: Covid-19 as Phenotype

Fakhare Alam , Obieda Ananbeh , Khalid Mahmood Malik * , Abdulrahman Al Odayani , Ibrahim Bin Hussain , Naoufel Kaabia , Amal Al Aidaroos

Posted Date: 19 January 2023

doi: 10.20944/preprints202301.0341.v1

Keywords: Deep Learning; COVID-19; Clinical Informatics; Machine Learning; Transformer; Association Mining

Article

Toward Predicting Length of Stay and Identification of Cohort Risk Factors Using Self-Attention Based Transformers and Association Mining: Covid-19 as Phenotype

Fakhare Alam [1], Obieda Ananbeh [1], Khalid Mahmood Malik [1,*], Abdulrahman Al Odayani [2], Ibrahim Bin Hussain [2], Naoufel Kaabia [2] and Amal Al Aidaroos [2]

[1] Department of Computer Science & Engineering, Oakland University, 115 Library Drive, Rochester, MI, 48309, USA; fakharealam@oakland.edu (F.A.); oananbeh@oakland.edu (O.A.)

[2] Infection Control Center of Excellence Prince Sultan Military Medical City; aalodayani@psmmc.med.sa (A.A.O.); ihussain@kfshrc.edu.sa (I.B.H.); nkaabia@psmmc.med.sa (N.K.); amalalaidaroos@gmail.com (A.A.A.)

* Correspondence: mahmood@oakland.edu; Tel.: +1-248-370-3542

Abstract: Predicting Length of Stay (LoS) and understanding its underlying factors is essential to minimize the risk of hospital-acquired conditions, improve financial, operational, and clinical outcomes, and to better manage future pandemics. The purpose of this study is to forecast patients' LoS using a deep learning model and analyze cohorts of risk factors minimizing or maximizing LoS. We employed various pre-processing techniques, SMOTE-N to balance data, and Tab-Transformer model to forecast LoS. Finally, Apriori algorithm was applied to analyze cohorts of risk factors influencing LoS at hospital. The Tab-Transformer outperformed the base Machine Learning models with an F1-score (.92), precision (.83), recall (.93), and accuracy (.73) for discharge dataset, and F1-score (.84), precision (.75), recall (.98), and accuracy (.77) for deceased dataset. The association mining algorithm was able to identify significant risk factors/indicators belonging to lab, X-Ray, and clinical data such as elevated LDH, and D-Dimer, lymphocytes count, and comorbidities such as hypertension and diabetes responsible for extending patients LoS. It also reveals what treatments has reduced the symptoms of COVID-19 patients leading to reduction in LoS particularly when no vaccines or medication such as Paxlovid were available.

Keywords: deep learning; COVID-19; clinical informatics; machine learning; transformer; association mining

1. Introduction

Worldwide, health systems infrastructure was severely strained by the rapid surge of patients infected with different coronavirus variants and many countries struggled to provide basic healthcare and timely services to patients [1]. Despite availability of Covid-19 vaccines, statistics shows that the hospitalization rate spikes globally in winter seasons during last two years [2] According to a study conducted in the US, inadequate critical care is associated with resource availability [3] and one hour of delay in services is associated with a 3% increase in patient mortality [4] An extended LoS is associated with a high risk of negative outcomes including adverse drug effects, hospital-acquired infections, inadequate nutritional levels, and many other complications [5] Inpatient care accounts for roughly a third of all healthcare spending in the U.S., with an average length of stay of 4.5 days and a daily cost of $10,400 [6]. Predicting Length of Stay (LoS) and understanding how to reduce it is one of the most critical factors for optimal usage of hospital infrastructure and medical resources during emergence of new infectious diseases. Thus, it improves financial, operational, and clinical outcomes by reducing costs for patients, such as facility expenses, supplies, and staffing. In addition, it minimizes the risk of hospital-acquired infections and reduces waiting time for patients. Thus, for precise resource management and utilization of the current infrastructure of the healthcare, a sophisticated approach is needed to predict the LoS, identify the cohort risk factors that leads to increased LoS, and understand what treatments can reduce the LoS.

For different diseases, LoS prediction was performed using statistical and conventional Machine Learning (ML) techniques such as Logistic Regressions (LRs), Random Forest Classifier (RFC), Decision Trees (DTs), etc. For example, Luo et al. [7] used LR and RFC to predict LoS in patients with pulmonary disease. Likewise, Dogu et al. [8] employed Artificial Neural Networks to predict LoS Chronic Obstructive Pulmonary Disease (COPD) patients while Kulkarni et al. [9] performed it using Multi-Layer Perceptron (MLP) for acute coronary syndrome patients. Likewise, to predict the ICU admission, mortality, and survivors' LoS for COVID -19, Dan et al. [10] created three ML prediction models without identifying cohorts of risk factors, and analysis was based only on univariate analysis. Lastly, Vekaria et al [11] developed statistical techniques such as truncation correction and survival bias to predict patients' LoS for COVID-19 patients but the employed data is not multimodal and suffers from quality issues such as missing timeline for discharged patients, and limitation of statistical model to recognize hidden patterns.

Deep Learning (DL) has proven capable of extracting complex, hidden correlations from data and achieved promising results when compared to existing ML methods. For instance, Zebin & Chaussalet [12] used an autoencoder deep neural network to categorize short stays (0–7 days) and long stays (>7 days) using Medical Information Mart for Intensive Care III (MIMIC III) dataset [13] but the dataset lacked multi-modalities and methodology does not analyze cohort of risk factors responsible for extended LoS. Likewise, Harerimana et al. [14] proposed an attention-based DL method to predict LoS and in-hospital mortality but this method also has limitations of using limited lab data and unavailability of radiological information. Rajkomar et al. [15] proposed a three-tier approach by combining three DL models to predict hospital readmission and patients LoS, based on data belonging to patients with varying diseases, without identifying the cohort of risk factors affecting patient LoS. These existing approaches only classify discharge or admit cases and predict the duration of the LoS using single modality data and are unable to provide the cohort of influencing factors that could increase or decrease patients' LoS, and most of them were not designed for predicting LoS for patients with infectious diseases.

The study design in this research focuses on detailed analysis of interaction and association between different risk factors and develop a framework for predicting the LoS of COVID-19 patients using state-of-the-art transformer-based architecture and identify groups of influencing factors occurring together and affecting LoS using pattern recognition techniques by mining multi-modal patient data such as lab data, x-ray data of Covid-19 patients.

The main contribution of this paper can be summarized as follows:

a. A state of art attention-based tab transformer model is presented to predict patients LoS using multiple modalities such as clinical features, patients' demographics data, and X-ray reports.
b. We present a framework where the result of machine learning based methods for LoS prediction can be analyzed with association mining rules and identify the cohort of risk factors affecting the LoS in hospitals.

The remainder of this paper is organized as follows. Section 2 describes the proposed method while section 3 presents experimental results and evaluation. Section 4 presents further discussion on the obtained results. Section 5 discusses limitations and future research directions followed by conclusion in Section 6.

2. Materials and Methods

The proposed framework for LoS prediction and identification of cohorts of risk factors includes three major components: Data Acquisition, Data Preparation, COVID-19 Risk Modeling. The detailed architecture and various subcomponents are shown in Figure 1. The main tasks in the data acquisition module include getting de-identified patients' data from Electronic Health Records (EHRs), identification of missing values with respect to main prognostic factors. The Data Preparation module includes data cleaning, quantization, balancing with respect to output (deceased, discharge) and LoS.

The COVID-19 risk modeling includes building and hyper tuning machine learning models and using association mining algorithms to identify cohorts of risk factors.

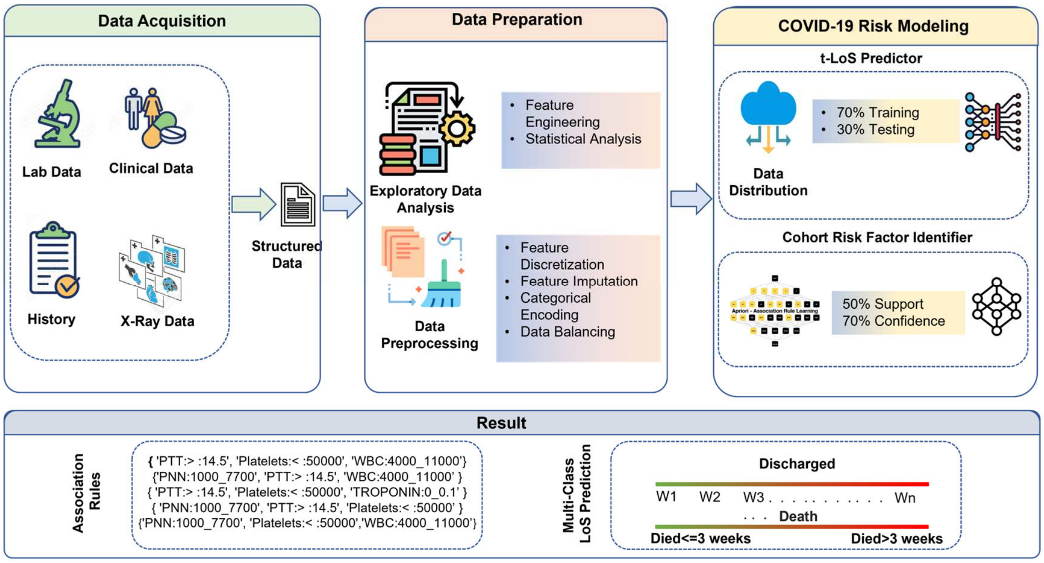
Data Acquisition
Lab Data
Clinical Data
History
X-Ray Data
Structured Data
Data Preparation
Exploratory Data Analysis
• Feature Engineering
• Statistical Analysis
Data Preprocessing
• Feature Discretization
• Feature Imputation
• Categrorical Encoding
• Data Balancing
COVID-19 Risk Modeling
t-LoS Predictor
Data Distribution
• 70% Training
• 30% Testing
Cohort Risk Factor Identifier
Apriori - Association Rule Learning
50% Support
70% Confidence
Result
Association Rules
{ 'PTT:> :14.5', 'Platelets:< :50000', 'WBC:4000_11000'}
{'PNN:1000_7700', 'PTT:> :14.5', 'WBC:4000_11000' }
{ 'PTT:> :14.5', 'Platelets:< :50000', 'TROPONIN:0_0.1' }
{ 'PNN:1000_7700', 'PTT:> :14.5', 'Platelets:< :50000' }
{'PNN:1000_7700', 'Platelets:< :50000','WBC:4000_11000'}
Multi-Class LoS Prediction
Discharged
W1 W2 W3 . . . . Wn
. . . Death
Died<=3 weeks
Died>3 weeks

Figure 1. COVID -19 LoS Risk Modeling Framework.

2.1 Data Acquisition

In this study, patients' data is obtained from Prince Sultan Military Medical City Infection Control Center of Excellence who were admitted between April 2020 to January 2021. There were 311 cases of confirmed COVID-19 infection, and after analyzing the data using prognostic factors, 308 cases were included in the analysis. This dataset includes 60 patients who died during treatment and 248 patients who were discharged. In total 89 features from three different categories: general information, X-ray and lab test are extracted. Table 1 shows the description of available features and frequencies with respect to different modalities data.

Table 1. Description of different modalities of data and available features.

Dataset Source	Description	Features Frequency
General	Contain general information such as demographic data (**gender**, **age**, and **ethnicity), epidemiological data (date of admission, date of death)** and **comorbidities** such as hypertension, diabetes, COPD etc.	68
Lab Data	Contain elements related to blood test such as WBC count, PNN, Lymphocyte's count, Hemoglobin, Platelets, Creatinine, ALT LDH, FERRITIN, D-DIMER, CRP, PROCALCITONIN, TROPONIN, Pro-BNP, PTT, Vitamin D and IL6	17
X-Ray Data	Contain elements related to X-Ray such as presence of **consolidation**, presence of **ground glass**, and opacities bilateral or unilateral	4

2.2. Data Preparation

In the data preparation stage, we performed basic data preprocessing such as binning, encoding, and initial Exploratory Data Analysis (EDA) by examining the distribution of attributes and summarizing the statistics of the data to uncover hidden patterns in the data. Table 2 shows the descriptive statistics about prominent features across different modalities of data. Also, during the EDA, we found that certain tests like aspartate aminotransferase, creatine, phosphokinase, and fibrinogen were not requested for some of the patients. Therefore, these features are not considered in modeling.

Table 2. Patients' main characteristics and descriptive details.

		Patient Characteristics	**Details: % Patients**
General	**Demographic**	Gender	Female: 49.7 %; Male: 50.3 %
		Age:	
		Mean	58.8 Years
		Median	60 Years
		IQR	26.7 Years
		Nationality	Egypt: 2%
			Filipino: 1.3%
			Iraq: .32 %
			Saudi Arabia: 95.7 %
			Sudan: .36 %
			United Kingdom: .32 %
	Comorbidities	Diabetes	69.2%
		Hypertension	64.3%
		Heart Ischemic	17.2%
		Heart Failure	5.0%
		Cardiomyopathies	1.3%
		COPD	2.0%
		Heart Failure	4.9%
		Lung Interstitial Disease	0.3%
		Bronchial Asthma	15.0%
		Cerebrovascular	4.2%
		Neurologic (Dementia)	4.2%
		Cirrhosis	1.3%
		HIV	0.0%
		Liver Disease	2.0%
		Shortness of Breath	85.7%
	Others	Psychiatric History	1.3%
		End Stage Renal	11.0%
		Hemodialysis	4.5%
		Cancer	6.0%
		Solid Organ Transplant	5.5%
		Hematopoietic Cell Transplant	0.0%
		Smoker	0.3%
		Pregnancy	5.0%
		Sick Cell	0.3%
		Obesity	5.5%
		Fever	55.0%

	Hemoptysis	1.0%
	Diarrhea	11.0%
	Cough	72.0%
	Headache	7.5%
	Abdominal Pain	8.0%
	Myalgia	11.0%
	Loss of Smell or Taste	8.0%
	Temperature	100.0%
	Respiratory Rate	13.6%
	Pulse	100.0%
	Nausea or vomiting	8.0%
	Diastolic BP	100.0%
	Systolic BP	100.0%
	Chest pain	4.0%
Lab Parameters	LDH	100.0%
	Glassgow	100.0%
	PaCO2	
	HCO3	
	PaO2	
	Ph	
	Lymphocytes	
	PaO2	
	WBC	
	ALT	
	PTT	
	D-Dimer	
	Platelets	
	WBC	
	Hemoglobin	
	CRP	
	Ferritin	
	AST	
	Pro BNP	
	PROCALCITONI	
	TROPONIN	
	Vitamin D	
	IL6	
	Blood Group	
	INR	
	Fibrinogen	
	PNN	
	Antiviral	
	Antibiotic	
	Anticoagulant	
Medications	Immunomodulators	80.0%
	Presence of Consolidation	98.0%
	Presence of Ground Glass Opacities	92.0%

	Bilateral or Unilateral	87.0%
X-Ray		72.0%

2.2.1. Natural Binning:

Data binning is a method used to minimize the effect of small observation errors. This is used to discretize continuous features and transform them into categorical features. We performed natural binning of all the continuous features such as age, Ph, PaO2, HCO3 etc. and then consulted with experts to adjust the binning boundaries. Binning introduces non-linearity and improves the performance of machine learning models by minimizing small observation error. Table 3 shows the optimized categorization for the continuous features.

Table 3. Optimized Binning of continuous features.

Patients' Feature	Optimized Binning Interval
Age	<=37 Years
	>=38 Years and <=55 Years
	>=56 Years and <=73 Years
	>=74 Years
pH	{<= 7.35; 7.35 - 7.45; >7.45}
PaO2	{<=80 mm Hg; >80 mm Hg}
PaCO2	{<= 35 mm Hg; 35 mmHg- 45mm Hg; > 45 mm Hg}
HCO3	{<=21 mEq/L); 21 mEq/L) - 27 mEq/L)}
Temperature	{<= 36 °C; 37.6 °C - 38.6 °C; > 38.6 °C}
Respiratory Rate	{<=12 bpm; 12 bpm - 20 bpm; 20 bpm - 28 bpm}
Pulse	{<= 79 bpm; 79 bpm - 95 bpm; 95 bpm -111 bpm; 111 bpm - 134 bpm; 134 beats per minute - 185 bpm}
Systolic Blood Pressure	{<= 90 mmHg; 90 mmHg - 130 mmHg}
Diastolic Blood Pressure	{<= 60 mmHg; 60 mmHg - 90 mmHg}
Glasgow	{<4; 4-8; 8-12; 12-14; > 14;}
WBC	{<=4000 /µL; 4000 /µL -11000 /µL; > 4000 /µL}
PNN	{<=500 mm3; 500 mm3 - 1000 mm3; 1000 mm3 - 7700 mm3;7700 mm3 - 15000 mm3}
Lymphocytes	{<=500 cells/µL; 500 cells/µL -1000 cells/µL; 1000 cells/µL -4000 cells/µL; >4000 cells/µL}
Hemoglobin	{<= 8 g/dl; 8 g/dl - 10 g/dl; 10 g/dl - 12 g/dl}
Platelets	{<=50000 /µL; 50000 /µL -150000/ µL; 150000 / µL - 450000 / µL}

Creatinine	{<= 59 mg/dL; 59 mg/dL - 104 mg/dL; 104 mg/dL - 250 mg/dL; 250 mg/dL - 500 mg/dL}
ALT	1 U/L -41 U/L;>41 U/L
LDH	{<= 135 IU/L; 135 IU/L - 225 IU/L}
FERRITIN	{<= 792; 792 -1976; 1976 - 4374; 4374 - 7627;7627 - 159000}
D_DIMER	{0 ng/mL-500ng/mL; >500ng/mL}
CRP	{<= 6 mg/L; 6 mg/L - 100 mg/L; >100 mg/L}
PROCALCITONIN	{<=0.25 ng/mL; 0.25 ng/mL - 0.5 ng/mL; > 0.5 ng/mL}
TROPONIN	{<=0.1 ng/mL;>0.1 ng/mL}
ProBNP	{<=12 pg/mL; 12 pg/mL -5 pg/mL; 5 pg/mL - 450 pg/mL}
PTT	{<= 11.5; 11.5 -14.5}
Vitamin D	{<=50 nmol/L; 50 nmol/L - 250 nmol/L}
IL6	{<=37.5 pg/ml; >37.5 pg/ml}

2.2.2. Encoding of Categorical Features

Our dataset comprises various categorical features including label values of (*0- No, 1-Yes*) such as diabetes and hypertension. It is essential that categorical features be transformed into numeric features before ML models can be trained effectively. We utilized an internal one-hot encoder; and converted various categorical features such to numerical features so that ML models can process it efficiently.

2.2.3. LoS Category Creation

Using the combined dataset from data acquisition step, we calculated the LoS using the admission and discharge or died timelines for all the patients and divided it into categories such as deceased within 3 weeks, deceased after 3 weeks, discharged within 1 weeks, discharged between 1 and 2 weeks, discharge between 2 and 3 weeks, discharged between 3 and 4 weeks, and finally discharged after 4 weeks from the date of admission. Table 4 shows the number of patients in each of these categories of LoS and patient's frequency.

Table 4. Data categorization and details of original and resampled data.

Classes	**LoS in Hospital**	**Patient Frequency Original**	**Patients Frequency After SMOTE -N**
Deceased	Less than or equal 3 weeks	36	36
	Greater than 3 weeks	24	36
Discharge	Less than or equal to 1 week	84	84
	Greater than 1 week and less than 2 weeks	79	84
	Greater than 2 weeks and less than 3 weeks	37	84

Greater than 3 weeks and less than 4 weeks	12	84
Greater than 4 weeks	36	84

2.2.4. Data Balance with Respect to LoS

The original dataset is imbalanced considering various LoS categories with-in discharged and deceased dataset. A balanced dataset is necessary to train a machine learning model to generate higher accuracy models and make unbiased decisions. The two primary approaches to make a balanced dataset out of an imbalanced dataset are under-sampling and oversampling. Given the limited number of patients in each categorized LoS, in this work we employ random oversampling using a variant of Synthetic Minority Oversampling Technique (SMOTE) called SMOTE-N [16]. This technique works well for categorical data such as diabetes, hypertension etc. and generates new instances from existing minority classes by taking samples of feature space for each target class and its nearest neighbors. This algorithm then generates new examples that combine features of the target case with features of its neighbors and increases the features available to each class and make the data more general. After balancing the data, the each LoS categories with in discharged (n=84) and deceased(n = 36) data contain equal records. Table 4 shows the original data and increased count of instances with-in each category after applying SMOTE-N.

2.3 COVID-19 Risk Modeling

The preprocessed and balanced data is used to develop and train a LoS Predictor model for both deceased, and discharged patients. At first, we developed a LoS Predictor model(t-LoSP) using state of the art transformer-based classifier followed by a Cohort Risk Factor Identifier (CRFI) to identify groups of risk factors affecting the patients LoS.

2.3.1 t-LoS Predictor

Tab-Transformer is an innovative recently developed deep tabular data model that can be used for both supervised and semi-supervised learning. Self-attention transformers are the foundation of the Tab-Transformers model. In the dataset, we have many categorical variables such as diabetes, hypertension, abnormal X-rays etc., the other available machine learning model such as neural network lacks the interaction and relationship between categorical variables in categorical embedding process. In the transformer-based architecture, the transformer layers convert categorical feature embeddings into strong contextual embeddings to improve prediction accuracy. Tab-Transformer architecture consists of a column embedding layer, a stack of N Transformer layers, and a multilayer perceptron. An individual transformer layer consists of a multi-head self-attention layer followed by a position-wise feed-forward layer. Figure 2 shows the detailed architecture of self-attention-based Tab-Transformer model and following are the steps for model execution:

Let x denotes input feature set and y multi-class target variable. Feature set x consist of both categorical ($X_{ca} = \{x_1, x_2, x_3 \ldots x_m\}$) and continuous variables (X_{co}). All categorical features are embedded into it to embedding space of dimension d using column embedding.

Let $e\varphi_j \in \mathrm{R}^{\mathrm{d}}$ for j $\in \{1,\cdot\cdot\cdot,\mathrm{m}\}$ be the embedding of the x_j feature, and $E\varphi(X_{ca}) = \{e\varphi_1\,(x_1),\cdot\cdot\cdot, e\varphi_m(x_m)\}$ are the embeddings for all categorical features.

The set of projected categorical embeddings $E\varphi(X_{ca})$ are input to the first transformer layer as shown in Figure 2. The output of the first transformer layer is sent to the next transformer layer and continues for N layer of transformer. The embedding output from individual layer is transformed into contextual embedding when resulted from the top layer of transformer, by consecutive aggregation of context from other embeddings. The sequence of transformer layers is denoted as a function f_{σ}. This function operates on parametric embeddings of categorical variables $\{e\varphi_1\,(x_1),\cdot\cdot\cdot, e\varphi_m(x_m)\}$ and results corresponding contextual embeddings $\{k_1, k_2 \cdot\cdot\cdot, k_m\}$ where $\mathrm{k_i} \in$ R d for i $\in \{1,\cdot\cdot\cdot,\mathrm{m}\}$. At the last the contextual embeddings obtained from transformer encoders

doi:10.20944/preprints202301.0341.v1

$\{k_1, k_2 \cdots, k_m\}$ are concatenated with the continuous features X_{co} to form a vector of dimension $(d \times m + c)$ and server as an input MLP classifier, denoted by h_ψ to compute target prediction variable y. Let J is the categorical cross-entropy for multiclass classification prediction task, we minimize the loss $J(x, y)$ to learn all the parameters of the Tab Transformer using gradient descent optimization method. The parameters of Tab Transformer include σ for Transformer layers, φ for column embedding, and ψ for the top MLP classifier.

$$J(x, y) \equiv H(g\psi(f_\sigma(E\varphi(x_{ca})), x_{co}), y) \qquad (1)$$

More information about tab-transformer architecture is available in [17]. The effectiveness of Tab-Transformer on multi-class datasets and particularly LoS prediction is unknown. In each deceased and discharged dataset, 70% of the data is used to train the model, and 30% to test its accuracy across multiclass prediction of LoS.

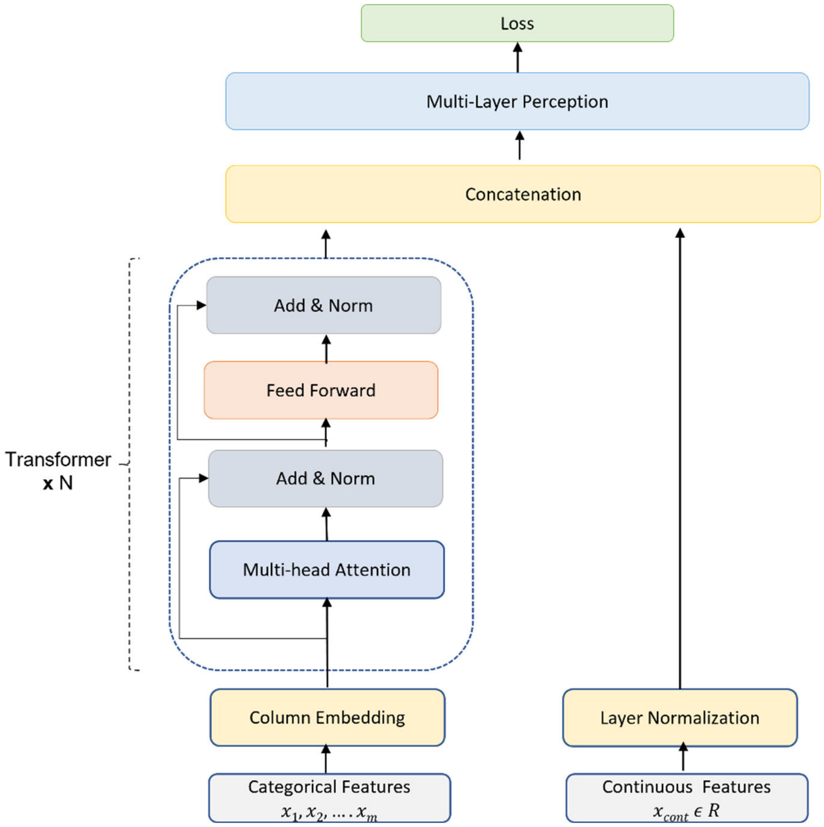

Figure 2. Tab Transformer architecture.

2.3.2. Cohort Risk Factors Identifier

The aim of CRFI is to identify the factor or combination of risk factors that had the greatest influence determining patients LoS in hospital. For this purpose, we employed the Apriori [18] that generates association rules by mining transactional data, which in our case are patient characteristics, symptoms, lab data, and X-ray features in each defined category of LoS. The association mining consists of the following four steps:

STEP: 1 Find all frequent itemset i.e., all patient characteristics appearing frequently together in the data with 50 % support and 70 % confidence.

$$Support_{(A)} = \frac{\text{Number of transactions in which A appears}}{\text{Total number of transactions}} \qquad (2)$$

$$Support_{(A \to B)} = \frac{\text{Number of transactions in which A and B appears together}}{\text{Total number of transactions}} \quad (3)$$

$$Confidence_{A \to B} = \frac{\text{Support}(A \to B)}{\text{Support (A)}} \quad (4)$$

where A, and B are itemset such as hypertension, diabetes, age ranges, and lab characteristics ranges etc. as defined in Table 3.

STEP: 2 Generate association rules from the aforesaid frequent itemset.

STEP: 3 Create a metric by calculating normalized harmonic mean of support and confidence using min-max scalar.

STEP: 4 Productionize the rule by selecting all the rules above threshold value (β =.7). This threshold value is decided after the rules are reviewed by the experts and considering the frequency of patients belonging to each rule.

3. Results and Evaluation

This section shows the performance of the proposed framework and highlights the important features in estimating the LoS of hospital patients. The implementation and experiments are performed on a 2.10 GHz Intel(R) Xeon(R) Platinum 8160 processor with Python programming environment.

3.1. COVID-19 Risk Model Results

We performed experiments with five ML models AdaBoost (AB), Decision Tree (DT), Gradient Boosting (GB), Logistic Regression (LR), Random Forest (RF), and a deep learning transformer-based model called Tab-Transformer (TabT) and used precision, recall, accuracy, and F1 score to compare the results. The hyper tuned tab-transformer model achieves the highest level of F1 score (discharged: .92; deceased: .84) compared to the base ML classifiers on both deceased and discharged dataset. Table 5 Shows comparative analysis of base machine learning models and tab-transformer model for LoS prediction on discharged and deceased dataset.

Table 5. Comparative Analysis of Tab-Transformer with baseline model on discharged and deceased dataset.

Classifiers	Discharge Dataset				Deceased Dataset			
	F1	**Accuracy**	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**
LR	0.74	0.73	0.77	0.74	0.68	0.68	0.7	0.73
RF	0.73	0.71	0.76	0.72	0.68	0.68	0.7	0.73
DT	0.65	0.65	0.68	0.65	0.62	0.64	0.64	0.66
AB	0.62	0.61	0.63	0.62	0.61	0.64	0.61	0.62
GB	0.54	0.52	0.61	0.53	0.50	0.5	0.6	0.6
TabT*	0.92	0.73	0.83	0.93	0.84	0.77	0.75	.98

3.2. CRFI Results

CRFI identifies cohort of risk factors associated with LoS and generates rules based on various patient characteristics. Table 6 illustrates the top sample rules for each category of LoS with in discharged and deceased patients' category. The complete rule set is publicly available to consume in GitHub repository [19].

Table 6. Top sample rules for discharged and deceased category LoS.

Dataset Type	LoS Category	Association Rules
Discharged Dataset	LoS ≤ 1 Week	{Anticoagulant, Cough, Antibiotics, Antiviral}
		{Cough, LDH> 225, Antibiotics, Antiviral}
		{Anticoagulant, SOB, Immunomodulators, LDH> 225, Antibiotics, Platelets< 50000}
		{PaO2 (0 to 80), Anticoagulant, SOB, LDH>225, Antibiotics}
	LoS >1 Week AND ≤ 2 Weeks	{Fever, DIMER (0 to 500), Immunomodulators, Antibiotics, Temperature (36 to 37.6)}
		{PaO2(0 to 80), CSA_Fever:0, Immunomodulators, LDH> 225, Antibiotics, Antiviral}
		{Anticoagulant, Fever, FERRITIN< 792, Immunomodulators, Glasgow> 14, Platelets<50000}
		{Anticoagulant, Fever, SOB, HTN, Glasgow> 14, Antiviral}
	LoS > 2 Weeks AND LoS ≤ 3 Weeks	{Fever, DIMER> 500, LDH> 225, Antiviral}
		{Anticoagulant, Fever, HTN, Diastolic BP (60 to 90), Antiviral}
		{Anticoagulant, Fever, HTN, Immunomodulators, Diastolic BP (60 to 90)}
		{CRP (6 to 100), Fever, LDH> 225, Antiviral}
	LoS > 3 Weeks AND LoS ≤ 4 Weeks	{Anticoagulant, Lymphocytes (1000 to 4000), Antibiotics, Respiratory Rate (20 to 28), PNN (1000 to 7700)}
		{Anticoagulant, HTN, Immunomodulators, Lymphocytes (1000 to 4000), Antibiotics, Respiratory Rate (20 to 28), Antiviral}
		{HTN, Immunomodulators, Lymphocytes (1000 to 4000), Antibiotics, Respiratory Rate (20 to 28), Antiviral}
		{Anticoagulant, Immunomodulators, Lymphocytes (1000 to4000), PNN (1000 to 7700)}
	LoS ≥ 4 Weeks	{Immunomodulators, Platelets< 50000, Antiviral, abnormal X-Ray}
		{Antibiotics, PTT> 14.5, Platelets< 50000}
		{Anticoagulant, Immunomodulators, Antiviral}
		{PNN (1000 to 7700), PTT> 14.5, Platelets< 50000, Antiviral}
Deceased dataset	LoS <=3 Weeks	{SOB, Antibiotics, PTT> 14.5, Platelets<50000, TROPONIN (0 to 0.1)}
		{SOB, Antibiotics, Glasgow> 14, PNN:1000_7700, Antiviral}
		{LDH> 225, Diastolic BP (60 to 90), Glasgow:> :14, Antiviral}
		{PaO2(0 to 80), Cough, Antibiotics, Glasgow> 14, Platelets< 50000}
	LoS >3 Weeks	{ALT (0 to 41), Diabetes, HTN, Immunomodulators, Antibiotics}
		{Ph (7.35 to 7.45), ALT (0 to 41), HTN, Immunomodulators, Antibiotics, Platelets < 50000}
		{ALT (0 to 41), SOB, HTN, Immunomodulators, Antibiotics, PTT> 14.5}
		{ALT (0 to 41), SOB, HTN, Immunomodulators, Glasgow> 14, Antiviral}

3.2.1. CRFI for Discharged Patients' Category

In the discharged patients' category, for LoS ≤ 1 week or Los ≤ 2 weeks, usage of anticoagulant, antibiotics and antiviral medications are important factors and indicates that timely intervention and dosage reduces LoS. For LoS ≤ 3 weeks, some of the important risk factors observed in the rules are elevated level of LDH (>225), D-Dimer (>500) and CRP (between 6 mg/L to 100 mg/L). Observed rules suggest that patients with abnormal values of these factors takes time to recover even if they provided with anticoagulant and antiviral medications. For LoS ≤ 4 weeks, the important risk factors observed are higher lymphocytes count (>1000 cells/µL), elevated PNN count (1000 -7000 mm3), comorbidities such as hypertension, higher respiratory rate (20-28 bps). The mining results on patients who stayed more than 4 weeks in the hospital shows less platelets count (<50000), abnormal X-ray, PTT>14.5, and higher PNN count. We found these patterns along with usage of antiviral, anticoagulant and antibiotics medications, which again suggest that abnormal values of above-mentioned risk factors increase LoS even if the medication are provided. The most affected (~40 %) age group for across all the category is (age >=56 Years and <=73 Years), closely followed by (age >=38 Years and <=55 Years). The observation suggests that age is an important factor in deciding the LoS.

3.2.2. CRFI for Deceased Patients' Category

In the deceased patients' category, Shortness of breath (SoB), low platelets count (<50000), diastolic blood pressure between (60 and 90), abnormal PTT (>14.5), higher LDH count (>225), low PaO2 (<80), and TROPONIN between (0 and .1) are the most critical factors for the patients who died within 3 weeks of admission to hospital. The critical risk factors for patients with LoS ≥ 3 weeks are ALT (0 - 41), SoB, comorbidities such as diabetes and hypertension, Glasgow (>14). The pattern suggests that abnormal values of these risk factors not only increase LoS but also increase the severity of COVID leading to patient death eventually. The most affected group is (age >=56 Years and <=73 Years) with 65%, 58.82%, 47.0%, and 56 % in rule 1, 2, 3, and 4, respectively. The younger generation (<=37) distribution across all the rules in deceased category is very minimal and for LoS>3, there is none from this age group.

4. Discussion

The current study aims to analyze cohorts of risk factors obtained from multimodal data by using state of art deep learning model tab transformer model and associating mining. The state of art tab transformer model shows excellent results on both deceased and discharged patients in predicting their LoS. The CRFI module analyzes the group of risk factors which extend the LoS in hospitals and result in either discharge or death. CRFI results show the identification of risk factors in cohorts can help in determining LoS and identifying criticality that leads to COVID severity.

Not much work has been done in determining LoS for COVID-19 patients using multimodal data. Examples and discussion of the prominent patterns are as follows:

- Age appears to be a strong risk factor for COVID-19 severity and its outcomes. Statsenko et al. [20] performed detailed analysis and concluded that elderly patients with COVID-19 are more likely to progress to severe disease. The result of CRFI within deceased category consists of rules with age >=56 Years and <=73 Years while other age category rules are not frequently observed and found to be insignificant. In addition, the mining results on patients who stayed in hospital between three and four weeks contains 25% of the rules with age>=73. These observations validate the fact the age is correlated with COVD-19 severity and a significant factor in deciding LoS.
- The detailed analysis on CRFI rules on the patients who stayed in hospital between 3 and 4 weeks showed that 43 % of the rules constitute either hypertension or diabetes, thus these comorbidities not only increase the LoS in hospitals but also leads to severe COVID leading to increased LoS in the hospital. Same was concluded by Adab et al., 2022 [21].

We also observe many key findings with respect to lab features such as LDH, dimmer and lymphocytes count. Few examples are outlined below:

- The elevated level of D-dimers is an indicator and major risk factor for thrombosis (blood clotting) and increases the risk of medication and monitoring for a longer time [22]. We observed that for the people who discharged between 3 and 4 weeks CRFI results with D-dimers shows that 18 % of the rules have D- dimers value more than 500 ng/mL FEU, thus increasing their LoS. In addition, mining results on patients who stayed more than 3 weeks and died elevated D-Dimer values is present in 41% of the rules. This is also validated by the fact that the people who discharge within two weeks, CRFI results show only 4.5 % of the rules has D-Dimer value more than 500 ng/mL FEU and elevated D-Dimers values are not found to be significant on the CFRI results of patients who stayed less than one week.
- LDH is another factor that has an elevated level of more than 225 units/L in 23% of the rules based on CRFI results of patients discharged from hospital between 3 and 4 weeks.
- Wagner et al. [23] concluded that lymphocytes count is one of the prognostic factors in determining COVID-19 illness, and our CRFI results for patients who died after spending more than 3 weeks in hospitals found that all the rules with lymphocytes consist of values are between

500 and 1000 while its values are between 1000 and 4000, 86 % of the time for patients who discharged within two weeks. This again validate the fact the lower lymphocytes count is critical in determining COVID severity and LoS.

- In the start of Covid-19, medical community tried many treatments without much evidence. It is important to understand what medications based on lessons learned could be useful to treat infections caused by new strains of viruses as viable epidemic response strategies. Our study shows that drug such as Hydroxychloroquine, Favipiravir, reduces the patients LoS. CRFI results on patients who stayed less than a week in hospital show 51% of the rules consist of antibiotic medications, while who discharged in less than 2 weeks, 52 % of the rules consist of antiviral medication. These analysis shows that the usage of antiviral and antibiotic medication has reduced patient's LoS.

5. Limitation and Future Directions

This is a single-institute retrospective cohort study to predict LoS. Using multi-center data, it is possible to further evaluate robustness of proposed framework. Future work will focus on acquisition of data for different ethnicities and countries. Our results also did not find every possible combination responsible for COVID severity and LoS, it only found the prominent one based on support and confidence of association mining model. There is a chance that some important cohort of risk factors are missed because they are not present in the data.

6. Conclusions

Predicting patients LoS is a complex task as it depends on many factors such as patient's history, existing comorbidities, and socio-economic factors. The evaluation shows that state-of-art tab transformer along with association rule mining can effectively predict the LoS and compare how combination of risk factors can help in predicting and reducing the LoS in the hospitals. The LoS of Covid-19 patients varies with respect to many factors such as severity of illness, comorbidities, other cohorts of risk factors. The proposed framework presented can be used for predicting LoS for infectious diseases along with many other critical diseases such as pulmonary disease, cardiovascular disease.

Author Contributions: Conceptualization, F.A. and K.M.; methodology, F.A and O.A.; software, F.A. and O.A.; Formal Analysis, F.A., O.A, K.M; validation, K.M., A.O., I.H., N.K. and A.A.; resources, K.M.; data curation, I.H., N.K.; writing—original draft preparation, F.A.; writing—review and editing, F.A., K.M.; visualization, F.A., O.A.; supervision, K.M.; project administration, K.M.; funding acquisition, K.M. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deputyship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number 959.

Institutional Review Board Statement: Imam Mohammad Ibn Saud Islamic University International Review Board; Reg HAPO-01-R001 approved this study.

Data Availability Statement: The datasets, libraries, and any supporting tools used or analyzed during the current work are accessible upon reasonable request from the corresponding author.

Acknowledgments: All the experimental protocols adhered to the accepted ethical guidelines and standards for human data usage and research study. We acknowledge the administrative and technical support of Imam Mohammad Ibn Saud Islamic University, Prince Sultan Military hospital, and Oakland University.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. W.H., et al. (2021). *Second round of the national pulse survey on continuity of essential health services during the COVID-19 pandemic: January-March 2021: interim report, 22 April 2021* (No. WHO/2019-nCoV/EHS_continuity/survey/2021.1). World Health Organization.
2. Mathieu, E. (2022, December 28). Coronavirus (COVID-19) Hospitalizations. Our World in Data. Retrieved December 28, 2022, from https://ourworldindata.org/covid-hospitalizations

doi:10.20944/preprints202301.0341.v1

3. Bravata, D. M., Perkins, A. J., Myers, L. J., Arling, G., Zhang, Y., Zillich, A. J., ... & Keyhani, S. (2021). Association of intensive care unit patient load and demand with mortality rates in US Department of Veterans Affairs hospitals during the COVID-19 pandemic. *JAMA network open*, *4*(1), e2034266-e2034266.
4. Churpek, M. M., Wendlandt, B., Zadravecz, F. J., Adhikari, R., Winslow, C., & Edelson, D. P. (2016). Association between intensive care unit transfer delay and hospital mortality: a multicenter investigation. *Journal of hospital medicine*, *11*(11), 757-762.
5. Resar, R., Nolan, K., Kaczynski, D., & Jensen, K. (2011). Using real-time demand capacity management to improve hospitalwide patient flow. *The Joint Commission Journal on Quality and Patient Safety*, *37*(5), 217-AP3.
6. Weiss, A. J., & Elixhauser, A. (2014). Overview of hospital stays in the United States, 2012: statistical brief# 180.
7. Luo, L., Lian, S., Feng, C., Huang, D., & Zhang, W. (2017, March). Data mining-based detection of rapid growth in length of stay on COPD patients. In *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)* (pp. 254-258). IEEE.
8. Dogu, E., Albayrak, Y. E., & Tuncay, E. (2021). Length of hospital stay prediction with an integrated approach of statistical-based fuzzy cognitive maps and artificial neural networks. *Medical & Biological Engineering & Computing*, *59*(3), 483-496.
9. Kulkarni, H., Thangam, M., & Amin, A. P. (2021). Artificial neural network-based prediction of prolonged length of stay and need for post-acute care in acute coronary syndrome patients undergoing percutaneous coronary intervention. *European Journal of Clinical Investigation*, *51*(3), e13406.
10. Dan, T., Li, Y., Zhu, Z., Chen, X., Quan, W., Hu, Y., ... & Cai, H. (2020, December). Machine learning to predict ICU admission, ICU mortality and survivors' length of stay among COVID-19 patients: toward optimal allocation of ICU resources. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 555-561). IEEE.
11. Vekaria, B., Overton, C., Wiśniowski, A., Ahmad, S., Aparicio-Castro, A., Curran-Sebastian, J., ... & Elliot, M. J. (2021). Hospital length of stay for COVID-19 patients: Data-driven methods for forward planning. *BMC Infectious Diseases*, *21*(1), 1-15.
12. Zebin, T., & Chaussalet, T. J. (2019, July). Design and implementation of a deep recurrent model for prediction of readmission in urgent care using electronic health records. In *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1-5). IEEE.
13. Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*(1), 1-9.
14. Harerimana, G., Kim, J. W., & Jang, B. (2021). A deep attention model to forecast the Length of Stay and the in-hospital mortality right on admission from ICD codes and demographic data. *Journal of Biomedical Informatics*, *118*, 103778.
15. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., ... & Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, *1*(1), 1-10.
16. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
17. Huang, X., Khetan, A., Cvitkovic, M., & Karnin, Z. (2020). Tabtransformer: Tabular data modeling using contextual embeddings. *arXiv preprint arXiv:2012.06678*.
18. Borgelt, C., & Kruse, R. (2002). Induction of association rules: Apriori implementation. In *Compstat* (pp. 395-400). Physica, Heidelberg.
19. GitHub – Covid19_research. (n.d.). Retrieved December 28, 2022, from https://github.com/smileslab/Covid19_research/tree/main/Association_Mining
20. Statsenko, Y., Al Zahmi, F., Habuza, T., Almansoori, T. M., Smetanina, D., Simiyu, G. L., ... & Al Koteesh, J. (2022). Impact of Age and Sex on COVID-19 Severity Assessed From Radiologic and Clinical Findings. *Frontiers in cellular and infection microbiology*, 1395.
21. Adab, P., Haroon, S., O'Hara, M. E., & Jordan, R. E. (2022). Comorbidities and covid-19. *bmj*, *377*.
22. Lehmann, A., Prosch, H., Zehetmayer, S., Gysan, M. R., Bernitzky, D., Vonbank, K., ... & Gompelmann, D. (2021). Impact of persistent D-dimer elevation following recovery from COVID-19. *PLoS One*, *16*(10), e0258351.
23. Wagner, J., DuPont, A., Larson, S., Cash, B., & Farooq, A. (2020). Absolute lymphocyte count is a prognostic marker in Covid-19: a retrospective cohort review. *International Journal of Laboratory Hematology*, *42*(6), 761-765.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.