應用深度學習模型於心肌梗塞致死併發症預測之探討

翁鼎傑　賴俊鳴　林祝興
東海大學資訊工程學系
jasonjason12570@gmail.com; cmlai@thu.edu.tw ; chlin@thu.edu.tw

摘要

一直以來醫學問題都是具有挑戰性的，並且醫學相關問題的解決需要謹慎與重要性兼具。在近期現代醫學當中雖然使用了現代治療方法，但仍有許多疾病的住院死亡率仍不樂觀。

本篇論文主要是針對心肌梗塞(Myocardial Infarction, MI)患者，在此次住院期間預測是否會有嚴重併發症產生[1]。在本文也利用特徵特性進行缺值填補(MICE)、運行各個演算法包含機器學習(Machine Learning)與深度學習(Deep Learning)來進行預測與比較評測，最終調適出達到94%準確率的深度學習模型。

關鍵詞：特徵醫學、深度學習、機器學習、Myocardial Infarction

Abstract

Medical problems have always been challenging, and the resolution of medical-related problems requires more caution and importance. In the recent modern medical, although modern treatment methods have been used, the hospital mortality rate of many diseases is still not optimistic.

This paper is mainly aimed at patients with myocardial infarction (Myocardial Infarction, MI), predicting whether there will be serious complications during the hospitalization period [1]. In this article, we also use feature characteristics to fill in missing values (MICE), various algorithms including machine learning and deep learning, for prediction and comparison evaluation, and finally build a 94% accuracy deep learning model.

Keywords: Medical, Deep Learning, Machine learning, Myocardial Infarction

1. 緒論

隨著醫療設備與衛生環境進步，傳統致命疾病已由傳染病相關改變為非傳染性疾病。在世界衛生組織(World Health Organization, WHO)公布之西元2019全球十大死因中，主要以三個大主題有關「心血管疾病（缺血性心臟病、中風）」、「呼吸系統疾病（慢性阻塞性肺病、下呼吸道感染）」、「新生兒疾病」，尤其以心血管疾病為死亡首因。

自西元2000年以來，缺血性心臟病是全世界最大的死亡主因，並且死亡人數是攀升最急遽的疾病，於2019以來與2018相比增加200萬人次，目前已達890萬人次死亡。心臟病發作與中風通常屬於急病重症，主要是由於堵塞血管導致無法正常流通血液或是腦部，常見原因則是因為過多脂肪堆積於血管與腦部的血管內壁。

心血管疾病內細分為許多疾病，以冠心病(Coronary Syndrome, CS)為首，尤其急性心肌梗塞(Acute Myocardial Infarction, AMI)不管是對非住院病患抑或是住院期間的病患更是會讓其處於一個高危狀態。在以往的病例當中，經過各個研究數據急性心肌梗塞之住院中死亡率平均落至4-20%[2]，並且伴隨著更多的嚴重併發症，包括心源性休克、肺水腫、心肌破裂、血栓栓塞與心室顫動等等，這些急性和亞急性病症通常是致死的主要原因結果。

2. 研究目的

針對心肌梗塞[3]病患，大約一半的患者會出現併發症，導致病情惡化甚至死亡，即使是經驗豐富的專家也不能總是預見這些並發症的發展。對此，預測心肌梗死的並發症以便及時採取必要的預防措施是一項重要任務。因此即便有了設備之進步以外，針對預防惡化與預測併發症發生是尤其重要[4]，本篇論文將利用機器學習與深度學習建構出良好的預測模型。

本論文之研究目的為構建出「針對心肌梗塞病患的致死併發症發生預測」模型[5]。利用進院生理數據與患者信息，將每個數據進行分析與拆解，將其作為模型的訓練與驗證，最後製作出適合的預測模型並且對比不同結構的模型。

3. 研究方法

本節介紹研究架構、數據集、數據處理、模型探討，這裡會詳細介紹本論文所使用的細節。

3.1 數據集

本次選用的資料集為UCI的公開資料集，收集於1992至1995年，並公開於西元2020年12月。這次的這個資料集主要是為了研究與探討有關於「Myocardial Infarction，MI」的相關病症研究，在這次的數據集當中總共有1700筆數據與124個特徵欄位，其中涵蓋的數據有包括說入院時間所有的生理參數數據、第一天、第二天與第三天結束時所疼痛次數與使用具有嗎啡作用的藥物統計，依照所述特徵集合而成的數據集。

在本篇論文所選的資料集當中，含有不少缺少的數據。在所有的特徵值中有110個特徵含有缺失值，大約是88%的欄位皆有缺失值。在個別特徵部分(Serum CPK content (KFK_BLOOD)、Heredity on CHD (IBS_NASL))在缺失數值部分高達90%以

上，並且有多個特徵值(Diastolic blood pressure according to Emergency Cardiology Team (D_AD_KBRIG)、Systolic blood pressure according to Emergency Cardiology Team (S_AD_KBRIG)、Use of NSAIDs by the Emergency Cardiology Team (NOT_NA_KB)、Use of lidocaine by the Emergency Cardiology Team (LID_KB)、Use of opioid drugs by the Emergency Cardiology Team (NA_KB))共計五個缺失率皆大於33%，其餘缺失部分狀態則較為輕微，**表 1**為前三大缺失數據之數值表格。

表 1 前三大缺失數據

特徵欄位	IBS_NASL	KFK_BLOOD	D_AD_KBRIG
缺失率	99.7%	95.7%	63.2%

在數據集中，我們的目標為構建出「針對心肌梗塞病患的致死併發症發生預測」模型，在採用的目標特徵欄位是「Lethal outcome (cause)」，0的標註為「存活」，其餘則是各項致死併發症，包括「cardiogenic shock」、「pulmonary edema」、「myocardial rupture」、「progress of congestive heart failure」、「thromboembolism」、「asystole」、「ventricular fibrillation」。**表2** 為 Lethal outcome (cause)欄位的原始數據分布，**圖1** 為此特徵欄位的數據分布。

表 2 Lethal outcome 欄位的原始數據分布

Lethal outcome (cause)	sample counts
Alive	1001
cardiogenic shock	70
pulmonary edema	40
myocardial rupture	20
progress of congestive heart failure	17
thromboembolism	15
asystole	10
ventricular fibrillation	10

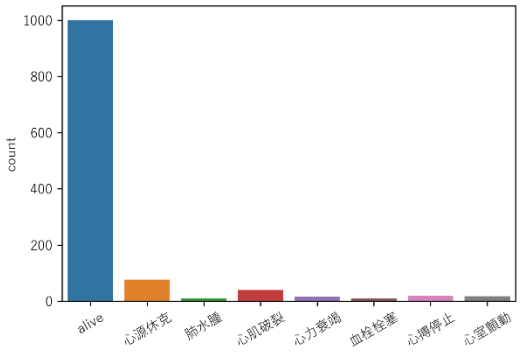

圖 1 目標特徵欄位數據分布

3.2 數據處理

數據缺失是數據集常有的問題，時常因為各種因素與環境，導致無法蒐集完全所有數值，因此這部分是數據處理第一個必須要解決的問題，而在這些丟失的數據當中，又有分為「完全隨機丟失 (MCAR, Missing Completely at Random)」、「非隨機丟失 (MNAR, Missing not at Random)」。這部分將會導致我們後續處理數據集的方式，包括說是否能夠直接丟棄抑或是需要將其作為另一種標示變量。以完全隨機丟失(MCAR, Missing Completely at Random)，數據集丟失的數據與其丟失之機率以及其他特徵值都完全無關；非隨機丟失(MNAR, Missing not at Random)，有些可能的情況以醫學部分來說，可能這個病人病情並不需要插管治療，則他的插管數據則會是缺失的狀態，因此這部分的特徵缺失則可以變成另一種的特徵狀態而去進行預測。所以說以這些情況與數據集，對於數值的有效識別是極為重要的。

而當數據不能以捨棄進行時，勢必會有缺失數據，而缺失數據會造成大量偏差使數據的處理與分析困難，於是「插補」被視為避免使用捨棄欄位的首要方法。目前有一些眾所周知的方法，包括說 hot deck imputation、cold deck imputation、mean imputation、regression imputation、stochastic imputation、multiple imputation…等等插補方法，在這篇論文中主要以三種插補方法進行探討。

在本篇論文中，首先針對捨棄動作部分，我們將較高比率之缺失值欄位進行捨棄動作，在這其中捨去的特徵值經觀察為隨機丟失（MAR, Missing at Random），因為過高的缺失若是進行強行補值動作，將會造成整體模型偏移，因此我們這邊將缺失率高於33%之特徵進行捨棄動作，並且我們將以116項特徵進行訓練與預測。**表3** 為本論文捨棄之特徵欄位，**表4** 為數據集可能的並發症（輸出）列，位於特徵欄位第 113-124 列。

表 3 本論文捨棄之特徵欄位

Droped Feature	Lost rate
Serum CPK content (KFK_BLOOD)	99.7%
Heredity on CHD (IBS_NASL)	95.7%
Diastolic blood pressure Emergency Cardiology Team (D_AD_KBRIG)	63.2%
Systolic blood pressure Emergency Cardiology Team (S_AD_KBRIG)	63.2%
Use of NSAIDs by the Emergency Cardiology Team (NOT_NA_KB)	40.3%
Use of lidocaine by the Emergency Cardiology Team (LID_KB)	39.8%
Use of opioid drugs by the Emergency Cardiology Team (NA_KB)	38.6%

表 4 數據集可能的並發症（輸出）列

Output Feature	Ratio	sample counts
Atrial fibrillation (FIBR_PREDS)	10.00%	170
Supraventricular tachycardia (PREDS_TAH)	1.18%	20
Ventricular tachycardia (JELUD_TAH)	2.47%	42
Ventricular fibrillation (FIBR_JELUD)	4.18%	71
Third-degree AV block (A_V_BLOK)	3.35%	57
Pulmonary edema (OTEK_LANC)	9.35%	159
Myocardial rupture (RAZRIV)	3.18%	54
Dressler syndrome (DRESSLER)	4.41%	75
Chronic heart failure (ZSN)	23.18%	394
Relapse of the myocardial infarction (REC_IM)	9.35%	159
Post-infarction angina (P_IM_STEN)	8.71%	148

接下來針對缺值補值部分[6]，缺失數據需要進行資料填補，而「插補」是用替換值去與缺失的數值做一個填補 。在本文中我們主要探討的是以下三種方法進行「均值插補 (mean imputation)」、「KNN (k-nearest neighbors)」、「多重插補 (Multiple Imputation)」。

並沒有最正確的，只有最適合的；對於插補的數據來說，能夠符合其原有分佈與數值才是最重要的。在這次所選的三種方法有各自的優點與缺點，並且有各自適合的使用環境[7]。以「均值插補 (mean imputation)」來說，這個方法的核心技術(1)是將其變量的均值替換掉任何缺失值，這個方法的好處是可以讓該變量的樣本均值不會因為插補而造成改變；反之其缺點會造成減弱其插補變量的相關性，造成多變量分析時發生問題。

$$\hat{y}_{mi} = b_{r0} + \sum_j b_{rj} z_{mij} + \hat{e}_{mi} \quad (1)$$

若是以「KNN (k-nearest neighbors)」來說，K-近鄰演算法的核心技術是將計算每一個樣本點的距離，相鄰之間距離的量測，估算缺值則使用相鄰觀測值的完整值來，簡單來說即是以下步驟 ；

1. 計算每一個點之間的距離
2. 利用 K 值決定鄰近之樣本點數目
3. 進行投票動作（若是在連續型資料當中，則是計算其所有樣本點的平均數）
4. 以投票結果決定類別 ； 以平均數作為最終結果

而距離部分該如何去計算，在存在缺失座標情況之下，KNN 通過忽略缺失值並且放大非缺失座標之權重來計算歐幾里德距離（Euclidean distance），(2)歐幾里德距離高維度情況計算公式是在高維度的情況下的計算公式:

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}. \quad (2)$$

KNN 是一個簡易易懂的模型，在多種類別分類時表現優異，然而缺點也顯而易見，所有變數變量的計算相當龐大，每個點之間的距離皆要計算，並且當資料樣本非平衡（Imbalanced Data）時, 會加大預測干擾則情況。

最後本論文最後使用的一個插補演算法-「多重插補 (Multiple Imputation)」[8]，在許多情況下我們會因為插補方法無法契合而造成有極多的雜訊，這個方法的核心是利用平均跨過多個插補數據集的結果來解決這個問題，遵循的三個步驟: 「插補」、「分析」、「合併」。本論文中採用的多重插補，是鏈式方程進行的多元插補(MICE，Multiple Imputation by Chained Equations)。在 MICE 方程底下，預設立場所有缺失的數值都是缺失的，利用其各變數之間缺失的特性，將缺失數值置入其他變數做回歸，並且重複循環直至所有的缺失值都被回歸模型的預測所替代。總而言之，MICE 使用了分而治之方法去估算數據集變量中的缺失情況，一次僅觀察一個遍量，使用數據集所有便量來預測該便量的缺失。這樣子的預測模型，基於回歸並且形式取決於關注變量之性質。

另外針對數據集結果不平衡的部分，我們對其做了特別的處理。在這份資料集當中，我們可以明確的知道說有7種結果**表 2**，其中一種是活下來，而其於則是因為各種併發症而死亡。所以對總體分類來說，可以大致分為兩類: 死亡與非死亡。在這樣子的狀況下，可以大幅提高資料不平衡的狀況 ，已這樣的處理可以使得整體的資料平衡達到原有的3倍左右，使得整體資料集之目的結果更加明顯。**圖 2** 為總資料集合併預測項目的數據集樣本，**表 5** 為總資料集合併預測輸出樣本數。

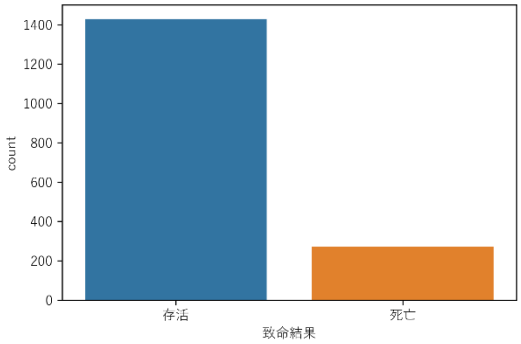

圖 2 總資料集合併預測輸出

表 5 總資料集合併預測輸出個數

存活個數	死亡個數
1429	271

但是好景不常，這樣的資料雖然已相較以往提高了不少，但是對於整個模型預測來說仍然處於資料不平衡的狀態，目前的資料平衡比為1:5，目標是希望盡可能的平衡整體狀態。在資料處理中，有方法能夠對資料集數量進行重構，透過個體的複製與刪除，對兩者拉近彼此距離，分別為升取樣(Upsampling)、降採樣(Downsampling)，這些是基於抽樣方法。

本論文選擇採用升取樣方法，除了在經過混淆來打亂樣本順序以外，僅將訓練集之樣本進行過採樣提升目標訓練樣本數，這樣子的做法既可以避免過多的特異性下降並且可以使得整體可預測樣本數量達到平衡，並且沒有經升取樣的驗證樣本可以表示目前的模型是可行的。在本論文中的數據集，將以7:3的比率來分配訓練集與測試集，**表 6** 為實際訓練集與測試集樣本數。

表 6 實際訓練集與測試集原有樣本數

訓練集總樣本數	測試集總樣本數
1190	510

圖 3 為經升取樣的訓練集，以樣本數看來特徵目標目前已達平衡。**表 7** 為實際預測輸出樣本數。

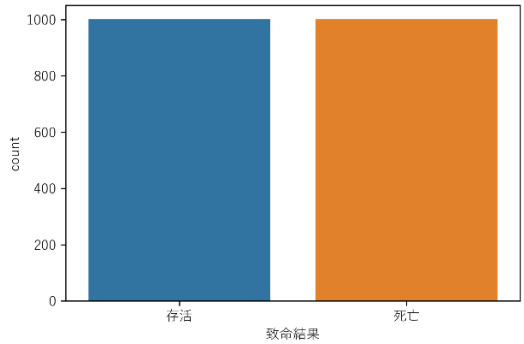

圖 3 升取樣後之訓練集

表 7 測試集合併預測輸出個數

存活個數	死亡個數
476	34

以這樣子的數據集進行訓練動作，可以大幅提升原始訓練集的訓練效果，讓整體較平衡。

3.3 模型探討

本論文所選之數據集特性與任務模型為分類型，故我們選擇使用 classification 類型模型，採用 DecisionTree[9]、Randomforest[10]、Logistic Regression[11]和近年來指標性模型 XGBoost(eXtreme Gradient Boosting)[12]進行比對與調整訓練。

在本論文中，我們主要以 XGBoost 作為主要探討的主題。XGBoost 是屬於 Boosting 算法的分支，主要核心架構皆是將許多弱決策樹集合在一起形成一個強的決策樹，以 Gradient Boosting 為基礎下去製作而成。保有了 Gradient Boosting 對每一個樹都有關聯性，透過不斷添加樹，與透過特徵分裂來添加一棵樹，這樣子的動作其實就是透過每一次去學習新的函數曲線，並且去修正預測的方差。而在這部分比較特別的是 XGBoost 並沒有採用所有特徵，而是基於隨機森林樹在生成之時隨機抽取特徵，故而每棵樹的生成不會拿到全部的特徵去進行決策動作。在基於以上特性，XGBoost 將可以提供有別於傳統算法更加穩定與優化的模型。

4. 研究結果

在傳統的機器學習模型當中，我們採取了一些比較常見的進行實作訓練與預測，我們也嘗試了近年指標性模型 XGBoost 來進行預測，可以看到以這些模型來預測的結果圖為**表 8**:

表 8 各模型準確率

Model	Precision
DecisionTree	82.7%
Logistic Regression	80.5%
Randomforest	88.2%
XGBoost	94%

可以明顯觀察到在以 DecisionTree 與 Logistic Regression 部分準確率仍較低迷，在 Randomforest 模型的調教下，準確率部分已達88.2%，並且在 XGBoost 模型下，準確率更是達到了94%。在這次的模型部分我們採用局部可解釋性模型–SHAP(Shapley Additive exPlanation)來輔助理解，利用這兩個套件去解釋模型對數據的分析與抉擇。

與傳統線性模型相比，XGBoost 擁有較佳的精度，但也因為其利用複雜的演算法與決策，在

對於數據與結果預測上無法直觀理解與解釋，在類似這樣的模型往往被稱做黑箱模型。所幸，在2017年中，[13]有篇論文中提出了 SHAP 之方法用以解釋與嘗試理解分析各種模型；在傳統我們用以優化預測能力即參考了特徵之重要性，直觀的反應各個特徵之重要性與影響，但是仍不能理解與判斷特徵與最終決策之關聯性。SHAP(Shapley Additive exPlanation)，每個預測樣本皆有模型所導出對應之預測量能，其即是每個特徵有對應的數值並用以來了解特徵對模型之決策影響力。

而在醫學領域中，每一個事物都關乎謹慎一詞，嚴謹的處理、合理的對待與結果，在前沿領域、預測模型中更是如此，事出總有原因，造成這些結果必有主因，我們更需要了解在這份資料集中，是什麼特徵影響了模型並且讓其做出如此決策。於是我們使用 SHAP 對我們的模型進行解釋性處理，**圖 4** 是在這個模型中最重要一些特徵值

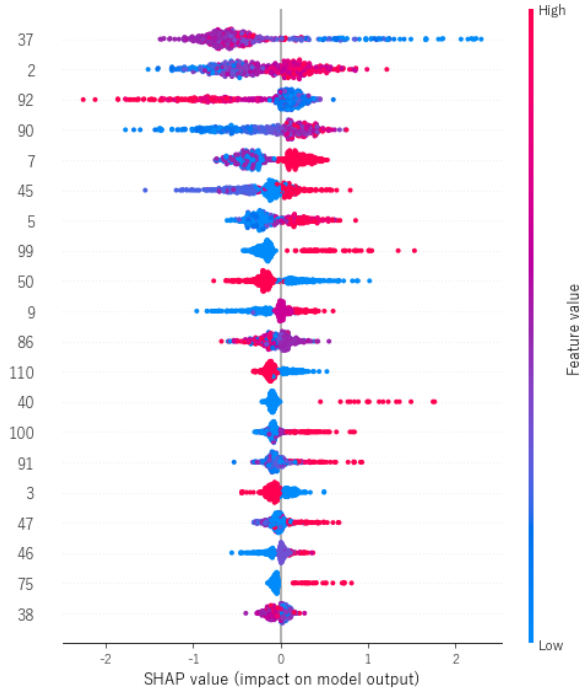

圖 4 重要特徵值與模型關聯性

表 9為前五大模型重要特徵值表

特徵欄位	關聯性
Systolic blood pressure according to intensive care unit	Negative
Age	Positive
Time elapsed from the beginning of the attack of CHD to the hospital	Negative
White blood cell count	Positive
Coronary heart disease (CHD) in recent weeks	Positive

在**表 9**中可以看到編號37與編號2的特徵值分別是重症監護室的收縮壓（S_AD_ORIT)與年齡(age)，在這次模型中的重要度占前幾名，並且在 Shap 分析中理解到重症監護室的收縮壓基本上與決策為負相關，收縮壓愈低越危險，反之年齡部分也可以很輕易地看出與結果正相關的情況，所以由 SHAP 分析出來的情況是年紀越大與收縮壓愈低擇越危險。

然而僅知道模型的情況是如何判定的仍不足以解釋單個病人的病情，在**圖 5** 中我們更可以利用 SHAP 進行單個病人的決策分析，利用 SHAP 更加了解我們的模型如何預測病人，是由哪些重要特質所決定與決策。

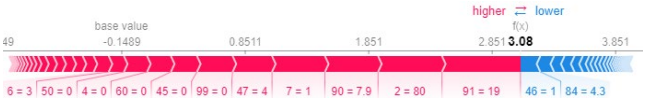

圖 5 單一病人 SHAP 分析

在這個病人中可以看到紅色為正相關特徵，藍色則為負相關特徵，並且主要影響的特徵都可以藉由 SHAP 進行分析出來。由此可見在所有數據中每個特徵值都有參與模型的決策，並由多個去做主要組成與判斷。

一個在醫療領域能夠實際應用的模型，必定要有相對準確，最重要的是能夠具有解釋性。透過 XGBoost 模型加上 SHAP 工具進行分析，使得這個題目能夠很好的被實驗與解釋，提供一個較為可行的目標與預測項目，並且這些經由 SHAP 分析出來的重要特徵，更有機會在實際醫療領域做為一個應用的預測特徵。

5. 結論

針對這次實驗各式模型演算法、數據分析工具與特徵分析工具 SHAP，我們提出一個應用於心肌梗塞致死併發症預測的模型，並且在有限的資料集中處理與分析，最後調適出了適合的模型並且給予了重要特徵分析。未來若是能夠得到更多相關資料集的樣本個數，想必可以讓這個預測將會更加準確與幫助相關議題研究的發展，借助 SHAP 工具也能讓大家一起揭開深度學習模型的黑盒子，正確的理解模型的決策過程並且可以在往後做修正與分析。

誌謝
本研究作者特別感謝 科技部贊助經費，使研究得以順利進行，計畫編號：MOST 109-2321-B-075A-001 -、MOST 107-2221-E-029-005-MY3 -。

參考文獻

[1] D. A. Rossiev, S. E. Golovenkin, V. A. Shulman and G. V. Matjushin, Neural networks for forecasting of myocardial infarction complications, The Second International Symposium on Neuroinformatics and Neurocomputers, Rostov on Don, Russia,1995, pp. 292-298, DOI: 10.1109/ISNINC.1995.480871

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Lu L, Liu M, Sun R, Zheng Y, Zhang P. Myocardial Infarction: Symptoms and Treatments. Cell Biochem Biophys. 2015 Jul;72(3):865-7. doi: 10.1007/s12013-015-0553-4. PMID: 25638347.

[4] Golovenkin, S.E., Bac, J., Chervov, A., Mirkes, E.M., Orlova, Y.V., Barillot, E., Gorban, A.N. and Zinovyev, A., 2020. Trajectories, bifurcations, and pseudo-time in large clinical datasets: applications to myocardial infarction and diabetes data. GigaScience, 9(11), p.giaa128., DOI:10.1093/gigascience/giaa128

[5] H. C. W. Hsiao, S. H. F. Chen and J. J. P. Tsai, "Deep Learning for Risk Analysis of Specific Cardiovascular Diseases Using Environmental Data and Outpatient Records," 2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE), 2016, pp. 369-372, doi: 10.1109/BIBE.2016.75.

[6] W. Kim, W. Cho, J. Choi, J. Kim, C. Park and J. Choo, "A Comparison of the Effects of Data Imputation Methods on Model Performance," 2019 21st International Conference on Advanced Communication Technology (ICACT), 2019, pp. 592-599, doi: 10.23919/ICACT.2019.8702000.

[7] T. Duy Le, R. Beuran and Y. Tan, "Comparison of the Most Influential Missing Data Imputation Algorithms for Healthcare," 2018 10th International Conference on Knowledge and Systems Engineering (KSE), 2018, pp. 247-251, doi: 10.1109/KSE.2018.8573344.

[8] Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. Int J Methods Psychiatr Res. 2011;20(1):40-49. doi:10.1002/mpr.329

[9] Fürnkranz J. (2011) Decision Tree. In: Sammut C., Webb G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_204

[10] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[11] Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232.

[12] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

[13] Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4 December 2017; pp. 4768–4777.