國立高雄科技大學
電子工程系博士班

博士論文

深度學習應用於生物醫學資訊學
Deep Learning in Biomedical Informatics

研究生：吳國銓 (Kuo-Chuan Wu)

指導教授：楊正宏 (Cheng-Hong Yang)

中華民國 一百零七 年 七 月

深度學習應用於生物醫學資訊學

Deep Learning in Biomedical Informatics

研 究 生： 吳國銓 (Kuo-Chuan Wu)
指導教授： 楊正宏 博士 (Dr. Cheng-Hong Yang)

國立高雄科技大學

電子工程系博士班

博士論文

A Thesis
Submitted to
Institute of Electronic Engineering
National Kaohsiung University of Science and Technology
in Partial Fulfillment of the Requirements
for the Degree of
Ph.D. of Engineering
in
Electronic Engineering

July 2018
Kaohsiung, Taiwan, Republic of China

**National Kaohsiung University of Applied Sciences is the predecessor of National Kaohsiung University of Science and Technology (renamed on Feb. 1, 2018)*

中華民國 一百零七 年 七 月

國立高雄科技大學(建工/燕巢)研究所學位論文考試審定書

本校 電子工程系 博士班

研究生 吳國銓 所提之論文

深度學習應用於生物醫學資訊學

合於 博士 資格水準，業經本委員會評審認可。

學位考試委員會

召集人 [signature] 簽章

委員 [signature] [signature]

[signature] [signature]

指導教授 [signature] 簽章

系所主管 [signature] 簽章

中華民國 107 年 7 月 24 日

Deep Learning in Biomedical Informatics

by

Kuo-Chuan Wu

A Thesis
Submitted to
Institute of Electronics Engineering
National Kaohsiung University of Science and Technology
in Partial Fulfillment of the Requirements
for the Degree of
Ph.D. of Engineering
in
Electronics Engineering

July 2018

Approved by:

Jin-Bor Chen _____ Hsueh-Wei Chang _____

Li-yeh Chuang _____ Wenyu Cj _____

_____ _____

Adviser: Cheng-Hong Yang _____

Chairman: _____

深度學習應用於生物醫學資訊學

學生：吳國銓　　指導教授：楊正宏 博士

國立高雄科技大學電子工程博士班

摘 要

生物醫學資訊學為目前跨領域的熱門領域，涵蓋電腦科學、生物資訊學、醫學資訊學三大學門。本論文旨在針對數值型的醫學病理特徵資料與生物 DNA 序列資料兩種不同類型資料，進行深度學習分析。根據不同類型的資料屬性，建構不同的深度學習模式，包含深度神經網路與卷積神經網路兩種模型，找出最適當的模型與參數進行分析。其一以腎臟科血液透析之醫學病理特徵資料，進行透析過程中病人發生透析低血壓可能的預測。其二，透過時間性的腎臟科病人血液資料，建構一個模型以預測未來資料的變化，利用 DNA 條碼序列進行物種辨識。近年來由於電腦科學的發展與突破，深度學習技術預期能成功的應用於生物醫學資訊學，本論文的實驗結果亦得到相當好的成效。即便如此，未來深度學習在各領域的應用仍存在著許多的挑戰。

關鍵字：生物醫學資訊學、深度學習、深度神經網路、透析低血壓、卷積神經網路、DNA 條碼

Deep Learning in Biomedical Informatics

Student: Kuo-Chuan Wu Advisor: Dr. Cheng-Hong Yang

Institute of Electronic Engineering
National Kaohsiung University of Science and Technology

ABSTRACT

Biomedical informatics is an interdisciplinary and popular research field that studies computer sciences, bioinformatics and medical informatics. This thesis focuses on two data types for deep learning, i.e. numerical data (quantitative variables) of medical pathological features and sequence data of biological DNA. Two deep learning models were constructed for different data and properties, including deep neural network (DNN) and convolutional neural network (CNN). The goal of model construction is finding the best and the most suitable model and hyper-parameters for analysis: firstly, predicting the patients of intradialytic hypotension occurrence during hemodialysis, and secondly, using sequence data of DNA barcodes is used for species classification. Recently, because of the computer sciences achieve performance breakthrough, deep learning technique is expected to apply to biomedical informatics successfully. In this thesis, the experimental results show that deep learning obtained outstanding performances. At present, although deep learning had completed many ad hoc applications and provided excellent results, that remains several potential challenges.

Keywords: biomedical informatics, deep learning, deep neural networks, intradialytic hypotension, convolutional neural networks, DNA barcode

Acknowledgement

國 銓 心 中 湧 百 感
立 誌 於 心 盡 叩 謝
高 朋 相 助 盈 眶 之
雄 厚 情 感 好 得 意
科 學 精 神 求 洋 溢
技 術 成 長 何 盼 於
大 話 之 家 雖 善 言
學 成 謝 詞 難 列 表

吳國銓謹誌於
國立高雄科技大學電子工程系
中華民國 一百零七年 七月

Contents

List of Figures

List of Tables

Chapter 1

Briefings in Biomedical Informatics

The definitions of biomedical informatics, medical informatics, clinical informatics or bioinformatics are extensively discussed but it is difficult to state with clarity. However, this thesis gathers as many literatures as possible about definition of biomedical informatics. In 2012, the term of biomedical informatics has been defined by American Medical Informatics Association (AMIA) [1] as shown follows, and other definitions are shown as **Table 1-1**.

> "*Biomedical informatics is the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, driven by efforts to improve human health*."

Table 1-1. List of definitions from textbooks and journals

Author (s)	Year	Description	Ref.
AMIA	2012	"*Biomedical informatics is the interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, driven by efforts to improve human health.*"	[1]
Bernstam, Smith and Johnson	2010	"*We propose that informatics is the science of information, where information is defined as data with meaning. Biomedical informatics is the science of information applied to, or studied in the context of biomedicine. Some, but not all of this information is also knowledge.*"	[2]
William Hersh	2009	"*It is the term used to describe the application of computers and technology in health care settings. Sometimes the term information and communications technology (ICT) is used when the use of HIT has a strong networking or communications component.*"	[3]
Shortliffe and Blois	2006	"*Biomedical informatics is the scientific field that deals with the storage, retrieval, sharing, and optimal use of biomedical information, data, and knowledge for problem solving and decision making.*"	[4]
Musen and van Bemmel	1997	"*biomedical informatics we develop and assess methods and systems for the acquisition, processing, and interpretation of patient data with the help of knowledge that is obtained in scientific research.*"	[5]
Greenes and Shortliffe	1990	"*the field that concerns itself with the cognitive, information processing, and communication tasks of medical practice, education, and research, including the information science and the technology to support these tasks.*"	[6]
Van Bemmel	1984	"*Medical Informatics comprises the theoretical and practical aspects of information processing and communication, based on knowledge and experience derived from processes in medicine and health care.*"	[7]

In decades, the biomedical informatics has been a newly arising branch of science, which includes translational bioinformatics, consumer health informatics, public health informatics, computational biology, clinical research informatics and clinical research informatics. The informatics field is a branch of information engineering which involves various disciplines, such as computer science, data science, management science, cognitive science, information system, information technology, statistics and mathematics [1]. With the era of big data has begun, discovering the knowledge from biomedical data has become big challenges and opportunities in biomedical informatics. Data-driven discovery science in biomedical informatics brings brand insights to help doctors and researchers with mining medical records, precision medicine and drug discovery. Various state-of-the-art technologies were proposed and improved successively, such as data mining, data sciences, machine learning and deep learning. Those technologies have been rapidly developing and made a major breakthrough in various fields. In biomedical informatics, several popular prediction models were proposed containing complex disease risk prediction, clinical effectiveness research, hospital admission prediction, and so on [1].

The application of information technology can be divided into six levels in biomedical informatics (see **Figure 1-1**), involving 1) communication and telematics, 2) storage and retrieval, 3) processing and automation, 4) diagnosis and decision making, 5) therapy and control, and 6) research and development. The complexity of levels increasing relies on human intelligence involvement [5]. A general framework of biology and medicine studies consists of system elements identifying, system modeling, systems understanding and systems controlling. Owing to new advances in technology, the artificial intelligence (AI) technique has been highly regarded in biomedical informatics, many literatures pointed out the AI solved complex problems. Integrating those systems and technologies is contributing to the progress of biology and medicine understanding [8].

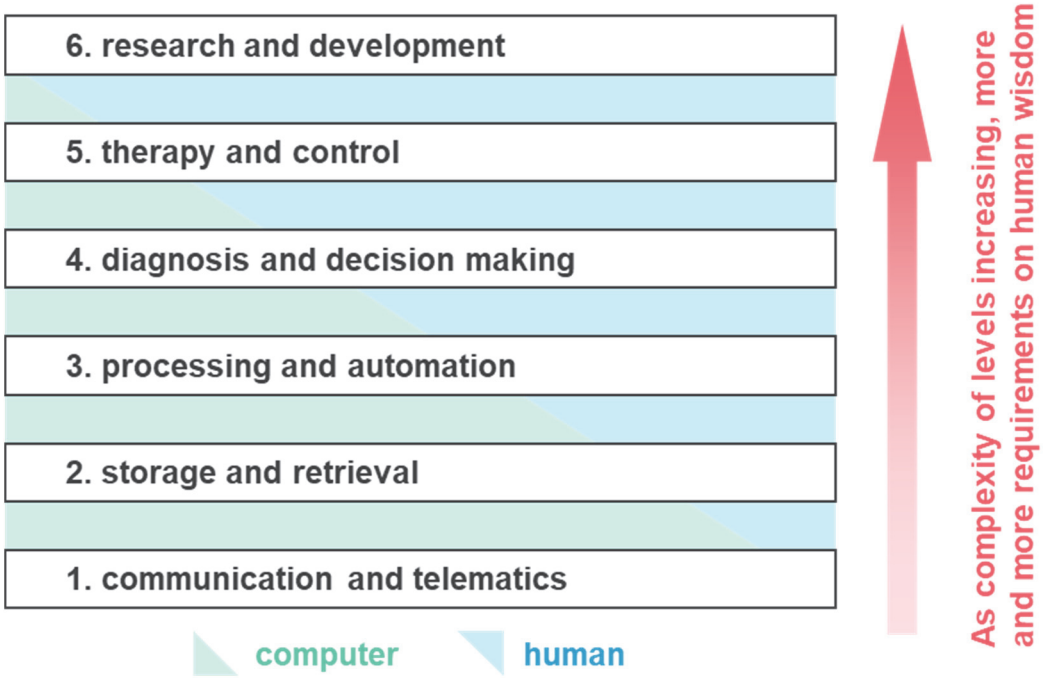


Figure 1-1. The application of information technologies in biomedical informatics

Source: modified from [5]

Chapter 2

Introduction of Deep Learning in Biomedical Informatics

Due to the mature development of cloud computing and big data, it has created a new trend in machine learning. Machine learning is a branch of artificial intelligence (AI) that endows the computers self-learning ability from human knowledges and insights. Machine learning technique can be roughly divided into three categories: supervised learning, unsupervised learning and reinforcement learning. In a word, the supervised learning is a technique that people design an algorithm to process the input data with labels to train a best model, and this best model can solve human tasks. The problems of classification and regression can be understood by two classical tasks. For example, given 10,000 pictures of dogs and cats to train machine (computer), query the machine what the new picture is, dog or cat? And then the machine can response the correct answer. Given the input data without labels, the unsupervised learning techniques can automatically

generalize a potential rule to explain the phenomenon and structure from the data. In brief, the machine will find out the potential rules automatically. For instance, based on the distribution and the similarity among data, the machine may find out the relation between buying diapers and also buying beers. Reinforcement learning consists of three major components: the agent, the environment and the actions. Give a task, the computer as an agent takes actions in an environment, those actions accompany with penalty. The intention of agent is gaining maximum reward. As an example of bowling competition, the machine will revise it progressively by the feedback of scores, and finally get a high score or proper result. In particular, deep learning is a part of machine learning which can be supervised and unsupervised learning. In some special tasks, reinforcement learning integrates with deep learning. In summary, classification (supervised learning) and clustering analysis (unsupervised learning) are two classical tasks in machine learning, different algorithms were proposed. This thesis organized the timeline of different algorithms in machine learning as shown in **Figure 2-1** and **Figure 2-2**.

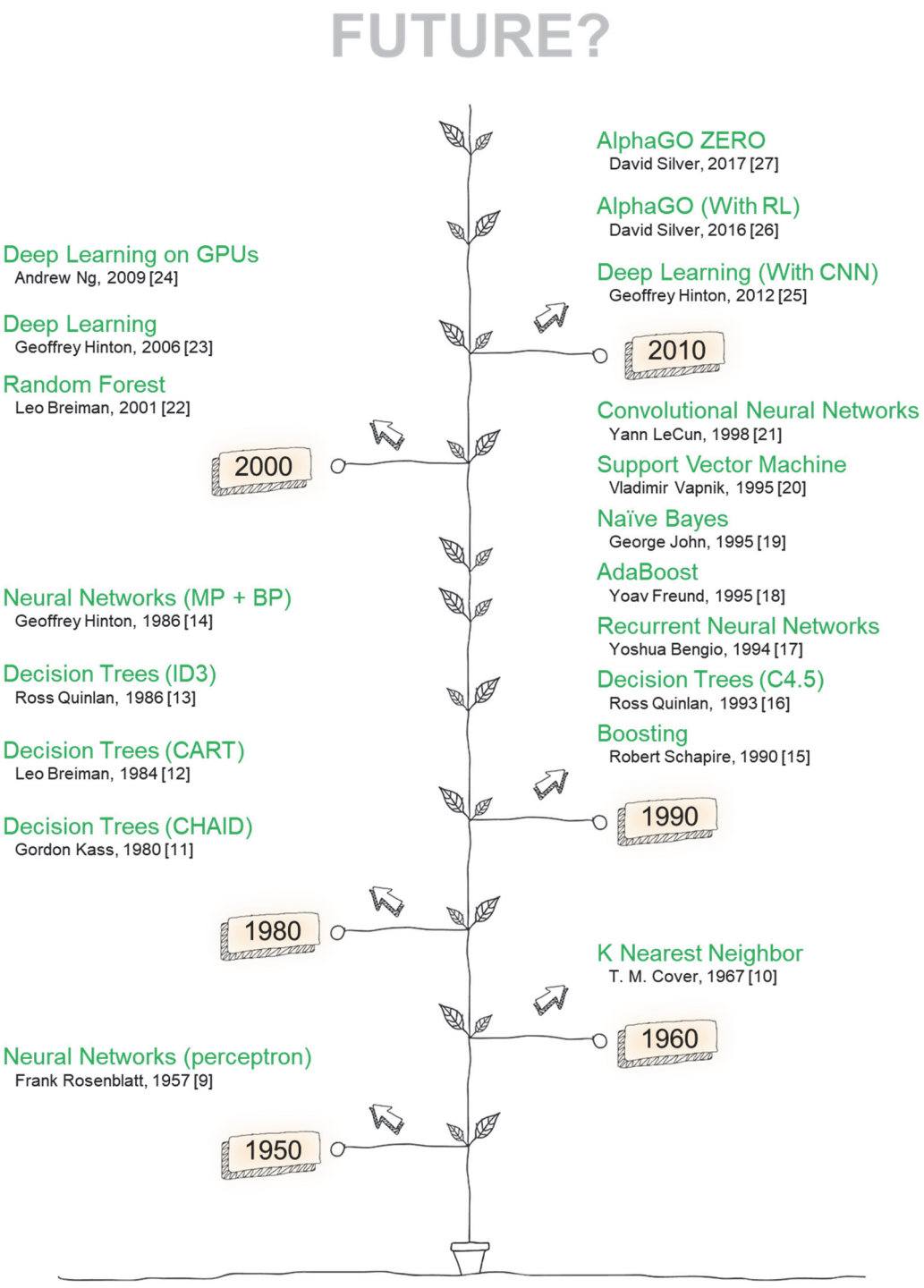


Figure 2-1. Timeline of classification in machine learning

*The picture material was modified from Freepik

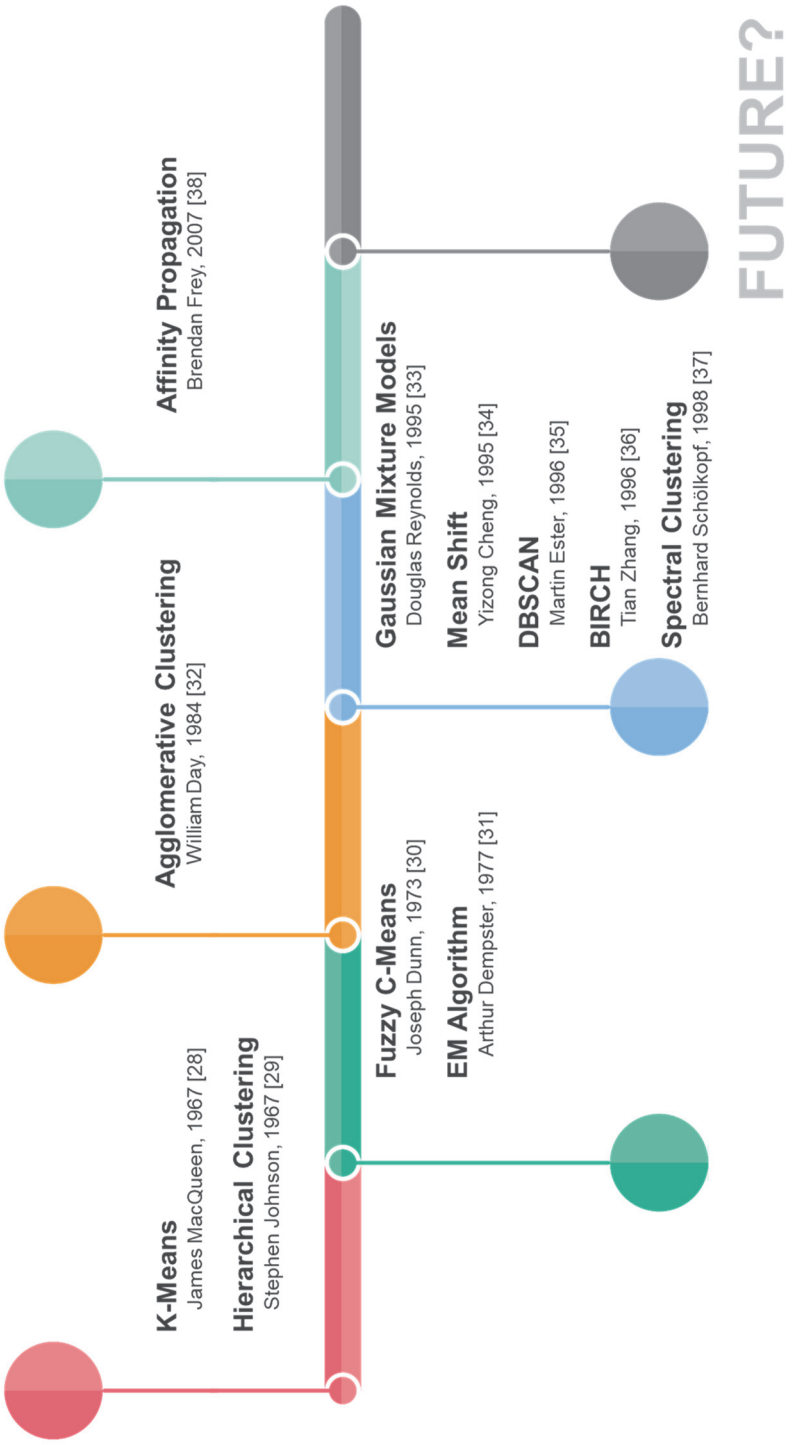


Figure 2-2. Timeline of clustering in machine learning

*The picture material was modified from Freepik

Main difference between deep learning and machine learning is machine learning system needs domain knowledge experts to transform the raw data into the handcrafted features as input. On the other hand, deep learning, instead of omitting the feature extraction in manual, automatically learns a complex model that maps input to output. Deep learning is an underlying mechanism method which derives competent prediction model without strong hypothesis. It computes and analyzes the intricate data, and automatically captures the texture of latent features in data which based on powerful parallel and distributed computing and sophisticated algorithms and that overcomes previous limitations and becomes to a state-of-the-art tool. A great deal of studies have demonstrated that deep learning accomplishes the outstanding performances in biomedical informatics [39-42].

Deep learning architectures can be categorized into serval types, including deep neural networks (DNNs) [14, 23, 43], convolutional neural networks (CNNs) [21, 43, 44], recurrent neural networks (RNNs) [17] [45], generative adversarial networks (GANs) [46], stacked auto-encoders (SAEs) [47]. Each architecture owns properties and advantages that is suitable for solving various tasks. More details of deep learning architectures were shown in Chapter 3 and Chapter 4.

This thesis presents two topic of solving prediction and classification tasks. Firstly, Chapter 3 introduces a model for intradialytic hypotension occurrence prediction during hemodialysis using DNN. Secondly, Chapter 4 presents a species classification with DNA barcode sequences using CNN, called DeepBarcode. Finally, Chapter 5 states the conclusion and future work.

Chapter 3

Prediction Model of Intradialytic Hypotension Occurrence during Hemodialysis Using Deep Learning

Intradialytic hypotension is a common problem during hemodialysis treatment. Several clinical variables have been authenticated for measurement during dialysis session. However, a report investigating the interaction between these variables by deep learning has not yet been presented. Our study aimed to investigate clinical factors associated with intradialytic hypotension by deep learning. A total of 279 participants undergoing regularly on an outpatient in a hospital-facilitated hemodialysis center were enrolled in March 2018. Associations between clinical factors and intradialytic hypotension were determined using linear regression method. The associated factors with $p < 0.2$ in a full-adjusted model were used in a deep neural network. A full-adjusted model indicated that intradialytic hypotension is positively associated with body mass index (Beta = 0.17, p = 0.028), hypertension comorbidity (Beta = 0.17, p = 0.008), and

ultrafiltration amount (% dry weight) (Beta = 0.31, $p < 0.001$), and is inversely associated with the ultrafiltration rate in a hemodialysis session (Beta = -0.30, $p = 0.001$). The 4-factor locus obtained by the deep neural network reached the maximum performance metrics evaluation (accuracy = 64.97 ± 0.94 (%); true positive rate = 87.97 ± 2.73 (%); positive predictive value = 66.74 ± 0.98 (%); Matthews correlation coefficient = 0.19 ± 0.03). The prediction model obtained by the deep learning resulted into a potential tool for the management of intradialytic hypotension.

3.1. Background

Intradialytic hypotension (IDH) is an uncommon event that occurs during a hemodialysis (HD) procedure. IDH is commonly defined as a decrease in the systolic blood pressure by ≥20 mmHg or in mean arterial pressure by ≥10 mmHg [48]. The pathophysiological mechanisms of IDH are complex. Two components have been discussed since past few years. First, an imbalance the between central hypovolemia and the adequacy of hemodynamic responses. In end-stage kidney disease, patients commonly manifest autonomic and baroreceptor dysfunction and disturbed cardiac function. Second, uncompensated plasma refilling occurs during the ultrafiltration procedure of HD. During HD, the ultrafiltration procedure removes the fluid from the vascular space and replaces the fluid in the interstitial space (plasma refilling). The rate of ultrafiltration during HD influences the rate of plasma refilling. When the amount of ultrafiltration exceeds the plasma refilling amount, IDH becomes an unavoidable event. Clinically, there are several diseases and circumstances apt to develop IDH during an HD procedure, namely diabetes mellitus, cardiac disease, autonomic neuropathy, severe liver disease, antihypertensive etc. [49].

Deep learning has been proved to be excellent for solving intricate problems and mathematical structures, and can be applied to a wide range of sciences, such as image/speech recognition, financial technology, computational biology and bioinformatics [43]. Deep learning is a complex computational architecture and is bio-inspired from human brains. The structure of brain neural networks consists of neurons as a layer of interconnected compute units. Several deep learning architectures have been categorized such as: deep neural networks (DNN), convolutional neural networks (CNN), recurrent neural networks (RNN), and other emergent or hybrid architectures [41]. The deep learning uses training data to discover underlying patterns and helps in constructing models and fitting the best model for prediction. Those models can be applied widely to bioinformatics, such as biomedical text recognition [50], biomedical imaging [51, 52], biomedical signal processing [53, 54], genomics [55, 56] and gene expression [57]. Kong and Yu [58] built a graph-embedded deep feedforward networks system for disease classification, and a biologically relevant feature selection using gene expression of RNA-seq data of kidney renal clear cell carcinoma. Sharma *et al*. [59] constructed a deep learning model to identify computed tomography images that estimated total kidney volume quantification for crucial autosomal dominant polycystic kidney disease

progression. Jackson *et al*. [60] developed a 3-dimensional automated image segmentation tool using deep learning, which detects renal segmentation of right and left kidney contours on non-contrast computed tomography images. In summary, deep learning is an emerging technique for renal study. Several literatures have revealed that deep learning has an advantage of better performance in achieving breakthroughs in various fields.

In the present study, we aimed to find the vulnerable variables, namely, demographics, comorbidities, laboratory parameters, vascular access parameters, reference values of HD machines during an event of IDH, components of hemodialyzers and HD drugs, by using deep neural network. The purpose was an attempt to find the significance of individual variables in the occurrence of IDH. Thus, optimal measures were expected to be applied in individual cases to prevent IDH.

3.2. Methods

3.2.1. Participants

The patients who were undergoing HD regularly on an outpatient basis at the Kaohsiung Chang Gung Memorial Hospital in Taiwan were enrolled for the investigation. Adult patients with IDH during the HD procedure were recruited. IDH was defined as the decrease in the systolic blood pressure by $\geq$20 mmHg during the HD procedure. The included patients were followed-up from 1st March 2018 to 31st March 2018. A total of 279 patients were eligible for inclusion in the IDH analysis. All the patients were undergoing HD every week. The protocol for the study was approved by the Committee of Human Research at Kaohsiung Chang Gung Memorial Hospital (documentation no: 201800595B0) for a retrospective review of the medical data.

3.2.2. Demographic data and clinical variables in hemodialysis session

Data collection was performed for the demographic information including age, gender, dry weight, body mass index (BMI), etiologies of end-stage kidney disease, comorbidities, and drug history. The notes related to the clinical variables of the HD

sessions were collected, including HD vintage, frequency of HD per week, duration of each HD session, ultrafiltration (UF) amount per hour (L/hour) and UF rate (L/hour) during each HD session, UF coefficient during the IDH event, blood flow rate (cc/min), vascular types, components of dialyzers, and occurrence time of IDH. The contents of the dialyzers were cellulose acetate (170G, FB210U), polysulfone (FXCor1000, FXCor60, PS-2.0W, PS-2.3W), polyether sulfone (EL-21H, EL-25H), and polymethylmethacrylate (BG2.1U). The dialysate flow was constant in all the HD sessions (500 ml/min). The temperature of the dialysate was 37 °C; the Ca concentration of the dialysate was 3.0 mEq/L and the bicarbonate concentration was 24 mEq/L.

3.2.3. Laboratory data

Baseline laboratory values for the blood analysis were measured in the midweek (on Wednesday or Thursday) via a venous port prior to the HD session, following an overnight fasting. The parameters included hemoglobin (Hb), albumin, blood urea nitrogen (BUN), creatinine (Cr), calcium (Ca), phosphate (P), sodium (Na), potassium (K), ferritin, and intact parathyroid hormone (iPTH) levels; the fractional removal of urea per dialysis treatment (Kt/V); urea reduction ratio (URR); normalized protein catabolic

rate (nPCR); and cardiothoracic ratio estimated by the chest x-ray examination. Detailed information about the measurements such as Kt/V urea, URR, nPCR, and cardiothoracic ratio have been described in our previous article [61].

3.2.4. Statistical analyses

The distribution of the continuous factors was summarized as mean and standard deviation or median and interquartile range, as appropriate; and the categorical factors were summarized in terms of frequency and percentage. Univariate and multivariate linear regression analyses were performed to demonstrate the interaction between the associated factors and intradialytic hypotension. The multivariate models, namely, a full-adjusted model was used. The full-adjusted model considered all the associated variables as covariates. A *p*-value less than 0.05 was considered as statistically significant. A cumulative incidence risk of intradialytic hypotension was visualized by using the Kaplan-Meier curve. All the statistical analyses were performed using the Stata software (StataCorp. 2009, Stata 11 Base Reference Manual, College Station, TX: Stata Press).

3.2.5. Data preprocessing

This study included 279 patients who were undergoing the hemodialysis treatment for 4 hours, which reported the occurrence time of intradialytic hypotension. We classified the patients into 2 groups; (1) those who reported the intradialytic hypotension occurrence time between 0 to 120 minutes, (2) those who reported the intradialytic hypotension occurrence time between 121 to 240 minutes. The baseline characteristics of the patients were pre-processed by using the standardization, which transferred the values of each factor from the center of the mean- and component-wise scale to the unit variance. Here, we used the data of all the factors to train each classifier model and test its performance; then, the p-value < 0.2 for the factors of the full-adjusted model were selected for a factor-interaction testing.

3.2.6. Deep neural network models

Deep neural network (DNN) is an artificial neural network architecture which consists of three parts, i.e. input layer, hidden layer and output layer. Within DNN model, multilayer computing units (called neurons) are highly interconnected with many hidden layers between input layer and output layer. In classification module, the conventional

neural networks are implemented fully connected feed-forward with neurons in multilayer. An illustration DNN is shown in **Figure 3-1**. Given n training data $X=\{x_1, x_2, \ldots, x_n\}$ with labels $Y=\{y_1, y_2, \ldots, y_n\}$ as input layer, an input x_i is mapped to next layer i.e. hidden layer h. Each neuron $h_j^{(\ell)}$ in ℓ^{th} hidden layer is computed as a input of non-linear system, describe as follows:

$$h_j^{(\ell)} = f\left(z_j^{(\ell)}\right), \text{ and } z_j^{(\ell)} = W_j^{\ell-1} \cdot h_j^{(\ell-1)} + b_j^{(\ell-1)} \quad (1)$$

where $W_j^{(\ell-1)}$ and $b_j^{(\ell-1)}$ are a weight and a bias in j^{th} column of matrix, respectively. $z_j^{(\ell)}$ is denoted as the outputs of the ℓ^{th} hidden layer, $z_j^{(0)}$ is defined as the input vector $\boldsymbol{X}$ (training data) to the network. These outputs are transferred by applying a nonlinear activation function f ( ). Here, we introduce a rectified linear unit (ReLU) for the activation function, define as follows:

$$\text{ReLU}(z) = \begin{cases} z & if\, z \geq 0 \\ 0 & if\, z < 0 \end{cases} \quad (2)$$

For multiclass classification, the last hidden layer is connected to the output layer into a probability of class k by using the *softmax* classifier, shown as follows:

$$o_j = \frac{\exp(z_j^{(\ell)})}{\sum_k \exp(z_k^{(\ell)})} \quad (3)$$

The best network parameters θ (i.e. weight $\boldsymbol{W}$ and bias $\boldsymbol{b}$) are optimized by

minimizing the cross-entropy loss function over training data which is defined as:

$$L(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left\|h_\theta\left(x^{(i)}\right) - y^{(i)}\right\|^2 \tag{4}$$

The gradient-based optimization algorithm for finding optimal parameters θ efficiently use backpropagation to find the gradient which performs as follows:

$$\theta = \theta - \eta\frac{\partial L(\theta)}{\partial \theta} \tag{5}$$

where η is the learning rate.

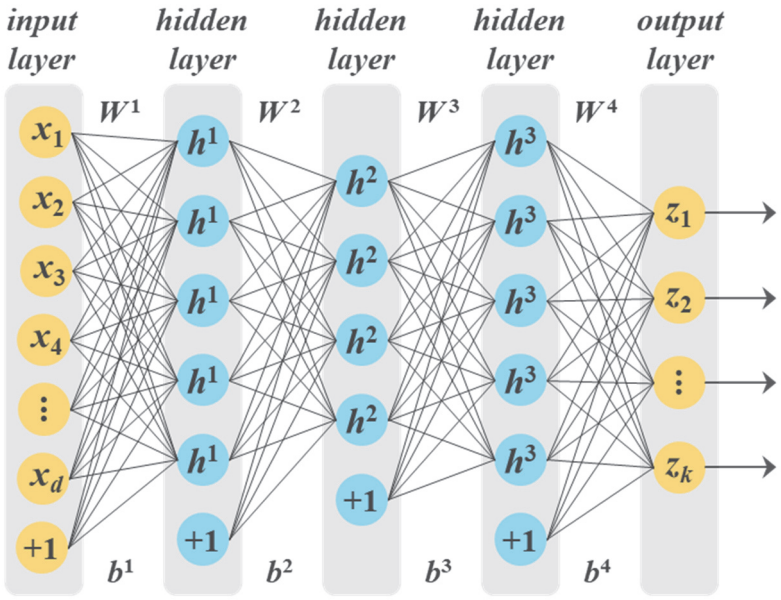


Figure 3-1. An illustration of a deep neural network with three hidden layers

3.2.7. Deep learning for the prediction of intradialytic hypotension

Figure 3-2 illustrates the progress of 5 steps in implementing a deep learning for the prediction of intradialytic hypotension. In step 1, a deep neural network (DNN) model was implemented in this study; we used a training set and a testing set with a *k*-fold cross-validation for training and testing the DNN model and factors selection model, respectively. The data were randomly divided into *k* equally-sized subsets. In step 2, *m* dimension (multifactor) of input data set for the prediction model was selected from the pool of all factors, where *n* factors have C_m^n possible combinations with *m* multifactor interaction. In step 3, the input data set was split into a training set and a validation set for the DNN prediction model creation using a holdout cross-validation method. In step 4, all the possible combinations of *m* multifactor interaction were trained and validated, and the best model with the validated identification accuracy was obtained. In step 5, the best prediction model of *m* multifactor interaction combinations was created, and then its performance was by the testing set. Finally, the *k*-fold cross-validation was repeated *k* times, the prediction accuracies were averaged, and a cross-validation consistency was obtained.

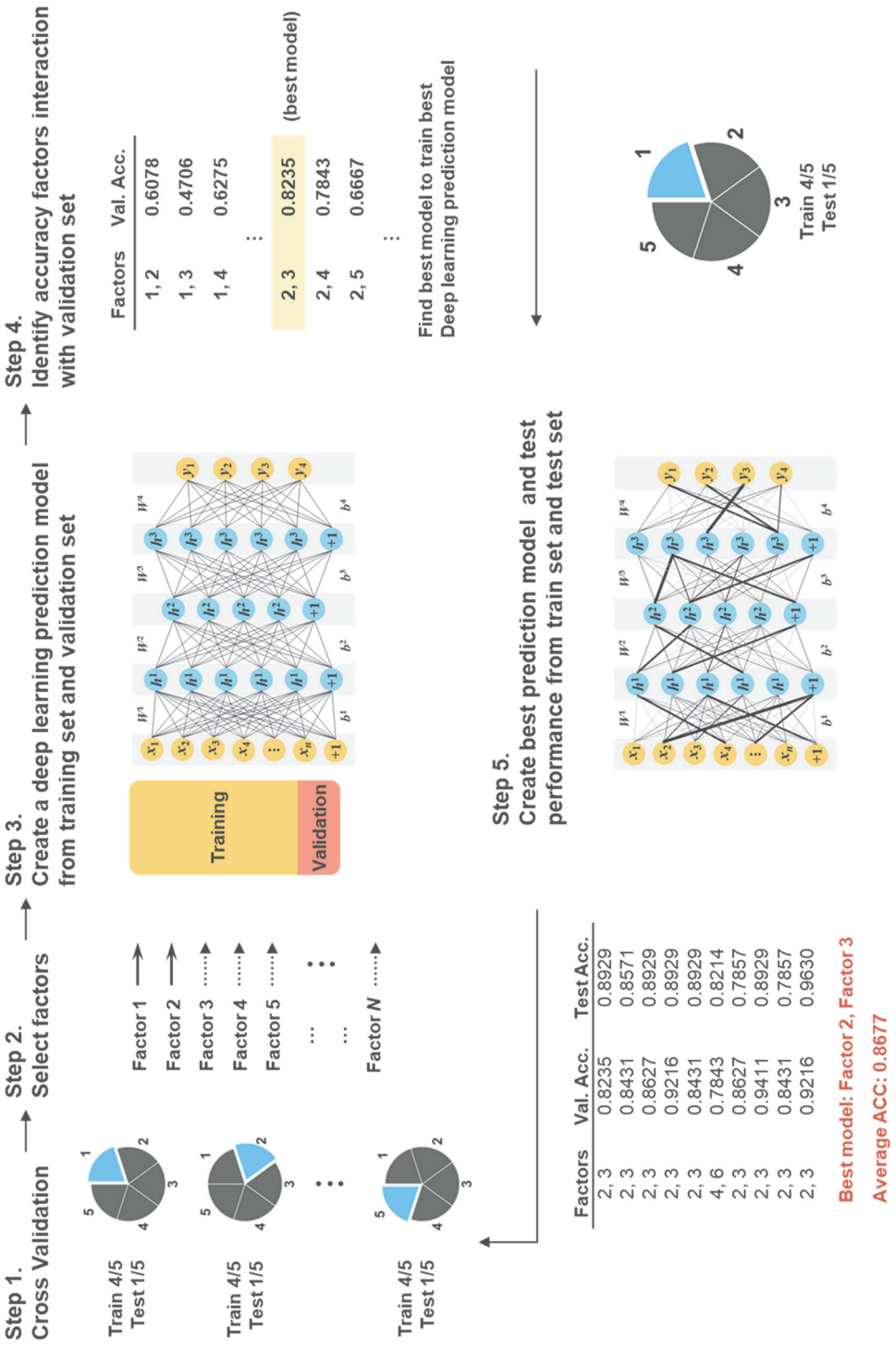


Figure 3-2. Overall factors interaction framework of dialysis hypotension prediction using deep learning

3.2.8. Stratified nested cross-validation

In overfitting problem, 2 experimental designs were used to acquire reliable and robust performance for overfitting avoidance [62]. Two stages of validation and testing were performed for both the experimental designs using the *k*-fold cross-validation and holdout cross-validation methods. The validation stage with the holdout cross-validation method was used to determine the best models of *m* multifactor interaction. The testing stage with the *k*-fold cross-validation method estimated the prediction performance by testing the previous best parameters and model obtained from the validation stage. The first experimental design used a stratified 5-fold cross-validation method; the data were randomly divided into 5 equal parts. One-fifth of the subjects in the testing stage and 4/5 of the subjects in the validation stage were classified into a nested stratified 5-fold cross-validation design. The second experimental design used the holdout cross-validation method; the data were randomly divided into 2 parts. The training set comprised of 80% of the subjects that trained the classifier model; whereas, 20% of the subjects formed the validation set, which validated the model performance. **Figure 3-3** shows an illustration of a simplified example of the 5-fold validation stage applied to 5 parts (P1, P2, …, P5) along with the best 2-factor model of all the 3 multi-factor combinations for some

classifiers. In the validation stage, the first loop including P1, P2, P3, P4 comprised the testing set; (1, 3) combination of 2-factor model showed the best performance (88% validation accuracy), and 90% test accuracy was obtained in the testing stage. Until finishing the testing of all the loops, an average accuracy of 88% was estimated as the final outcome.

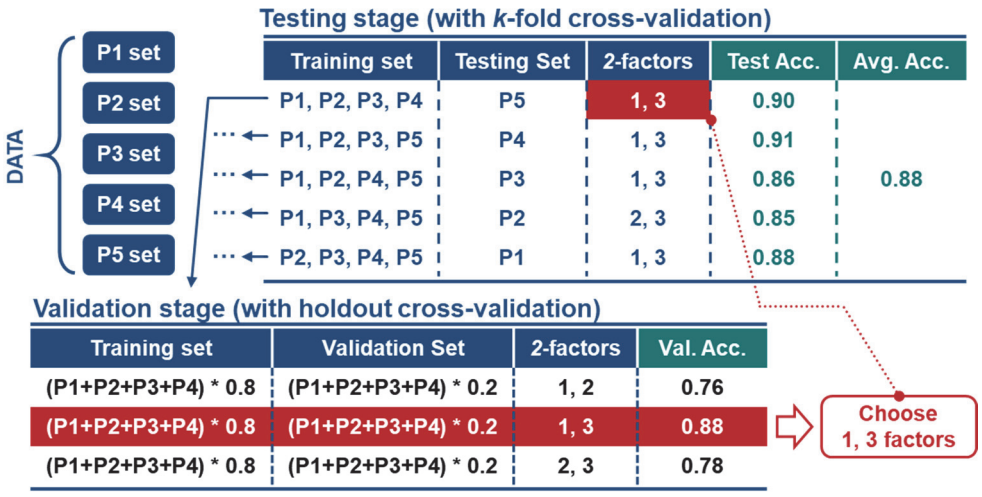


Figure 3-3. Illustration of simplified example for two experimental design

3.2.9. Architecture of the proposed DNN model

The DNN model was trained to classify the clinical characteristics of the patients on hemodialysis into 2 classes using a *softmax* cross-entropy objective function. In this study, we set 5 hidden layers (*h)* in the model using a rectified linear unit (ReLU) activation function ($\boldsymbol{H} = \{h^{(1)}, h^{(2)}, h^{(3)}, h^{(4)}, h^{(5)}\}$). Number of neurons were set as 128, 64, 64, 64, and 128 in $h^{(1)}$ to $h^{(5)}$ hidden layers, respectively. Adam optimizer was used to train the models after several epochs with various hyper-parameters, such as, an initial learning rate: 0.01, maximum number of epochs: 100, and the batch size on the cross-validation set: 10. All the DNN models were implemented using a Keras toolkit (https://keras.io/, based on Tensorflow [63] libraries in Python language).

3.3. Results

3.3.1. Association between the IDH and clinical variables

Totally, 279 patients with the mean aged of 63.04 years were enrolled; the clinical characteristics of the patients are summarized in **Table 3-1**. There were 136 (48.7%) male and 143 (51.3%) female patients. The cumulative incidence risk of IDH while undergoing HD treatment is summarized using a cumulative curve, as shown in **Figure 3-4**. The intradialytic hypotension occurred more likely after 30 minutes of undergoing HD treatment.

Table 3-2 demonstrates the association between the IDH and clinical variables. The univariate analysis indicated that IDH is positively correlated with the hypertension comorbidity (Beta = 0.15, $P = 0.012$) and calcium level (Beta = 0.13, $P = 0.033$), and is inversely correlated with the cardiothoracic ratio (Beta = -0.13, $P = 0.032$). The full-adjusted multivariate model indicated that IDH is positively correlated with the BMI (Beta = 0.17, $P = 0.028$), hypertension comorbidity (Beta = 0.17, $P = 0.008$), and UF amount (% dry weight, Beta = 0.31, $P < 0.001$), and is inversely correlated with the UF rate (Beta = -0.3, $P = 0.001$).

Table 3-1. Baseline characteristics

Variables	Total (n = 279)	
	Mean	SD
Age (year)	63.04	±11.8
BMI (kg/m^2)	23.39	±4.11
Body weight (dry) (kg)	61.10	±12.99
Gender (n, %)		
Male	136	48.7
Female	143	51.3
Etiology in ESRD (n, %)		
A (Renal Parenchymal Diseases)	80	28.7
B (Systemic Diseases)	168	60.2
E (Hereditary Diseases)	2	0.7
F (Other Causes of Renal Failure)	3	1.1
G (Renal Failure, Cause Unknown)	26	9.3
Comorbidity (n, %)		
DM	111	39.8
Hypertension	44	15.8
Others (Cirrhosis, CAD)	6	2.2
Antihypertensive	63	22.6
Dialyzer (n, %)		
170G	29	10.4
BG2.1U	45	16.1
EL-21H	53	19.0
EL-25H	52	18.6
FB-210U	56	20.1
FXCor1000	22	7.9
FXCor60	6	2.2
PS-2.0W	14	5.0
PS-2.3W	2	0.7
Vascular access (n, %)		
Fistula	231	82.8
Gore-Tex	23	8.2
Catheter	25	9.0

Abbreviations: BMI, body mass index; ESRD, end-stage renal disease; DM, diabetes mellitus; CAD, coronary artery disease.

Figure 3-1. *Continued*.

Variables	Total (n = 279)	
	Mean	SD
Blood flow rate (ml/min)	250	230-270
UF coefficient (ml/h/mmHg)	76	41.7-82
UF amount (% dry weight)	3.56	±1.58
UF rate (L/hour)	0.74	±0.28
Mean arterial pressure (MAP) in beginning (mmHg)	99.67	89.67-112.33
Mean arterial pressure (MAP) in hypotension (mmHg)	84.67	72.33-94.67
SBP beginning (mmHg)	153	137-173
SBP in hypotension (mmHg)	124	109-140
SBP reduction from beginning (mmHg)	31.04	±11.23
Occurrence time of hypotension (minutes)	102	63-141
Laboratory measurements		
Hb (g/dL)	10.53	±1.08
Albumin (g/dL)	3.92	±0.33
Ca (mg/dL)	9.51	±0.89
P (mg/dL)	5.29	±1.48
K (meq/L)	4.75	±0.75
Na (meq/L)	135.11	±3.35
Ferritin (ng/mL)	362.4	213.7-538
BUN (mg/dL)	73	58-88
Cr (mg/dL)	10.50	±2.71
Kt/V urea	1.45	±0.52
iPTH (pg/mL)	276.9	116.6-604.6
nPCR (g/kg/day)	1.17	±0.45
URR (%)	0.70	±0.3
Cardiothoracic ratio (%)	0.52	±0.08

Abbreviations: UF: ultrafiltration; Hb, hemoglobin; Ca, calcium; P, phosphate; K, potassium; Na, sodium; BUN, blood urea nitrogen; Cr, creatinine; iPTH, intact parathyroid hormone; nPCR, normalized protein catabolic rate; URR, urea reduction ratio

Table 3-2. Linear regression analysis

Variables	Univariate		Full-adjusted Model	
	Beta	*p*	Beta	*p*
Age (year)	-0.08	0.159	-0.13	0.080
BMI (kg/m^2)	0.04	0.488	0.17	**0.028**
Gender, Male	-0.10	0.106	-0.11	0.158
Comorbidity				
DM	0.02	0.745	0.06	0.356
Hypertension	0.15	**0.012**	0.17	**0.008**
Antihypertensive	0.03	0.631	-0.02	0.700
Vascular access, Gore-Tex vs Fistula	-0.001	0.984	0.06	0.321
Vascular access, Catheter vs Fistula	-0.06	0.341	0.023	0.725
UF coefficient (ml/h/mmHg)	0.10	0.089	0.101	0.184
Blood flow rate (mL/min)	0.02	0.781	-0.1	0.207
UF amount (% dry weight)	0.08	0.159	0.31	**<0.001**
UF rate (L/hr)	-0.04	0.474	-0.3	**0.001**
Mean arterial pressure (MAP) in beginning	0.05	0.449	-0.01	0.893
Laboratory measurements				
Hb (g/dL)	0.002	0.978	-0.02	0.738
Albumin (g/dL)	-0.04	0.523	-0.08	0.316
Ca (mg/dL)	0.13	**0.033**	0.12	0.062
P (mg/dL)	-0.01	0.847	-0.07	0.285
K (meq/L)	-0.01	0.916	0.04	0.551
Na (meq/L)	-0.03	0.566	-0.05	0.474
Ferritin (ng/mL)	-0.08	0.193	-0.032	0.618
BUN (mg/dL)	-0.04	0.558	-0.06	0.366
Cr (mg/dL)	0.07	0.218	0.02	0.833
Kt/V urea	0.00	0.961	-0.05	0.598
iPTH (pg/mL)	0.07	0.24	0.07	0.293
URR (%)	0.03	0.646	0.02	0.769
Cardiothoracic ratio (%)	-0.13	**0.032**	-0.12	0.077

p-values for categorical variables were calculated by χ^2 test, and continuous variables were calculated by independent two samples. $p < 0.05$ are highlighted in bold for each factor.

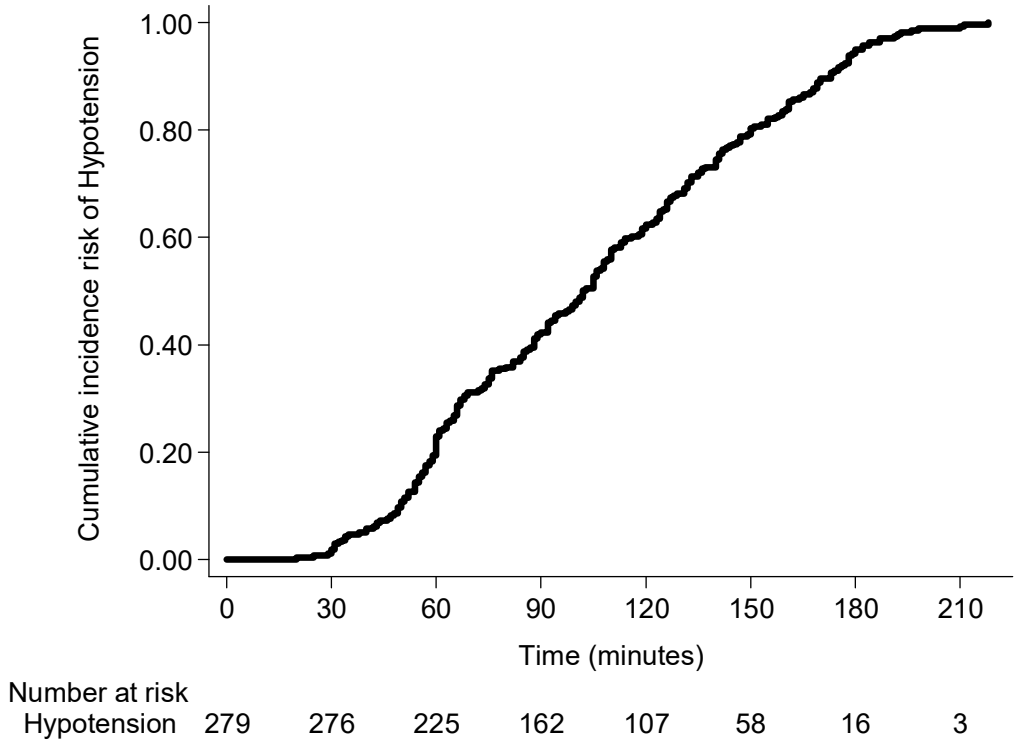


Figure 3-4. Cumulative incidence risk of hypotension in HD patients during receiving hemodialysis treatment

3.3.2. Performance comparison

Several general classifiers were implemented as the prediction models, referring the scikit-learn library v0.19.1 [64] in Python language, namely, support vector machine (SVM), artificial neural network (ANN), random forest (RF), decision tree (DT), K-nearest neighbor (KNN), and naïve Bayes (NB). The performance metrics of binary classification task were estimated from the complete feature set of each classifier by performing the 5-fold cross-validation on the dataset. The average receiver operating characteristic (ROC) curve and area under the curve (AUC) score were plotted, as shown in **Figure 3-5**. We calculated the ROC by using a *sklearn.metrics* package in the scikit-learn library v0.19.1 [64]. The experimental results revealed that, our DNN model achieved the highest performance (AUC = 64.90±3.11) than others. The ROC for our DNN model (ROC curve drawn in red line) is approximately 10% higher than the sub-best curve, which has been achieved by the RF model (ROC curve drawn in dashed purple line). This reports the comparative analysis showing that our DNN model outperformed the other models in superior performance and outstanding robustness in quantitative assessment.

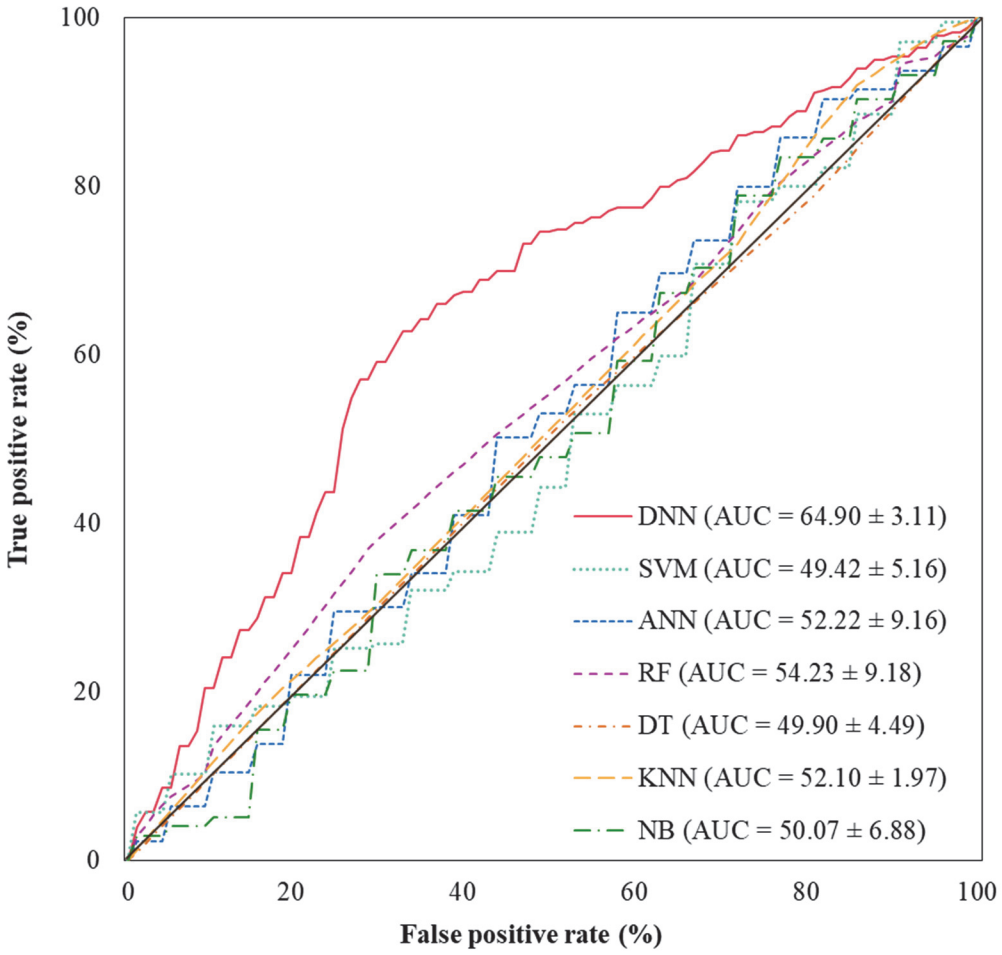


Figure 3-5. ROC curves of the classifier models for dialysis hypotension occurrence prediction on five-fold cross-validation

3.3.3. Factors interaction

We estimated the performance in terms of accuracy (ACC), sensitivity (e.g. true positive rate, TPR), precision (e.g. positive predictive value, PPV), and Matthews correlation coefficient (MCC), which were calculated as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. Particularly, ACC = (TP + TN) / (TP + FP + TN + FN); TPF = TP / (TP + FN); PPV = TP / (TP + FP); and MCC = [(TP × TN) – (FN × FP)] / $\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}$.

Table 3-3 summarizes the average ACC, TPR, PPV, and MCC values, obtained by using our DNN model after the analysis of the clinical dataset of the patients on hemodialysis, investigating the factor-factor interaction and performing the 5-fold cross-validation up to 100 evaluation cycles. The 4-factor locus revealed the maximum values after the performance metrics evaluation (ACC = 64.97 ± 0.94 (%); TPR = 87.97 ± 2.73 (%); PPV = 66.74 ± 0.98 (%); and MCC = 0.19 ± 0.03). The multi-factor reduction resulted into superior performance. Although, the accuracies obtained were below 70%, the TPRs were obtained above 85%. Generally, a high TPR (i.e. sensitivity) is important where the test is used to predict the occurrence of IDH during early hemodialysis treatments. To summarize, the leading factors showing a cumulative effect in the

occurrence of IDH during HD were the hypertension comorbidity, UF coefficient, UF rate, and UF amount.

Table 3-3. Summary of results for multifactor interacting

Num. of factors	Best model	ACC (%)	TPR (%)	PPV (%)	MCC
2	A-D	63.24 ± 1.18	91.54 ± 2.61	64.48 ± 0.70	0.13 ± 0.04
3	C-D-H	62.35 ± 1.96	87.55 ± 3.02	64.71 ± 1.30	0.11 ± 0.06
4	D-E-F-G	64.97 ± 0.94	87.97 ± 2.73	66.74 ± 0.98	0.19 ± 0.03
5	B-D-E-F-G	63.72 ± 1.32	85.15 ± 2.79	66.39 ± 1.06	0.17 ± 0.04
6	B-D-E-F-G-I	63.41 ± 1.27	83.98 ± 2.66	66.43 ± 1.08	0.16 ± 0.04
7	A-B-C-E-F-G-H	62.96 ± 2.00	80.96 ± 3.51	66.82 ± 1.20	0.16 ± 0.04

Abbreviations: A, age; B, BMI; C, Gender; D, comorbidity of hypertension; E, ultrafiltration coefficient; F, ultrafiltration amount; G, ultrafiltration rate; H, Ca; I, Cardiothoracic ratio; ACC, accuracy; TPR, true positive rate; PPV, positive predictive value; MCC, Matthews correlation coefficient.

3.3.4. Cumulative risk effect based on interaction model

The patients are divided into low-and high-risk group according to their IDH occurrence. The high-risk group is defined as the patients with IDH occurrence time less than 120 minutes, while the low-risk group is defined as the patients with IDH occurrence time greater than 120 minutes.

The scatter plot for all factors include in 2- to 7-factor interaction model is summarized as a matrix in Figure 3-6 to Figure 3-11. According to the matrices, the IDH time show not clear correlation with the individual factor. Thus, we used the ROC analysis to determine the best cut-off point of each factor for IDH low-and high-risk group in order to reduce the dimension complexity, and enable us to determine the characteristic which is highly associated with IDH during hemodialysis. The ROC curve for each individual factor is presents in **Figure 3-12** to **Figure 3-18**. The cut-off point with the highest value of addition in specificity and sensitivity is considered as the best cut-off point for each factor. The range of AUC for individual factor is ranged from 51.81 to 57.77, represents a low estimation accuracy for IDH low-and high-risk group.

The summary of 2- to 7-factor combinations associated with high-low risk for IDH is presents in **Figure 3-19** to **Figure 3-24**. In 2-factor interaction model, patients without

hypotension in both age groups obtained higher risk in rapid IDH occurrence. In 3-factor interaction model, the patients with lower calcium level (< 9.55 mg/dl) and without hypertension, female patients with higher calcium level (≥ 9.55 mg/dl) and without hypertension, male patients with higher calcium level (≥ 9.55 mg/dl) and hypertension obtained higher risk in rapid IDH occurrence. In 4-factor Interaction model, patients with lower UF amount (< 2.46 % dry weight) and without hypertension, patients with lower UF amount (< 2.46 % dry weight) and hypertension and higher UF coefficient (≥ 87.5 ml/h/mmHg) and higher UF rate (≥ 0.66 L/hour), patients with higher UF amount (≥ 2.46 % dry weight) and without hypertension and higher UF rate (≥ 0.66 L/hour) and lower UF coefficient (≥ 87.5 ml/h/mmHg) were considered obtained higher risk in rapid IDH occurrence.

The cumulative risk performance in 2- to 7-factor is determine using ROC analysis and the results is summarized in **Figure 3-25** to **Figure 3-30**. The interaction model with cumulative risk consideration showed a better results compare to the individual factor risk estimation. The AUC showed a rising trend from 2- to 6-factor interaction model, and slightly reduced in 7-factor interaction model. We found the 4- to 6-factor interaction model include four same factors (UF amount, UF rate, UF coefficient and hypertension),

while the 7-factor interaction model have no include the hypertension. The decreasing of AUC in 7-factor interaction model might related with the exclusion of hypertension which is considered as an important factor in rapid IDH occurrence.

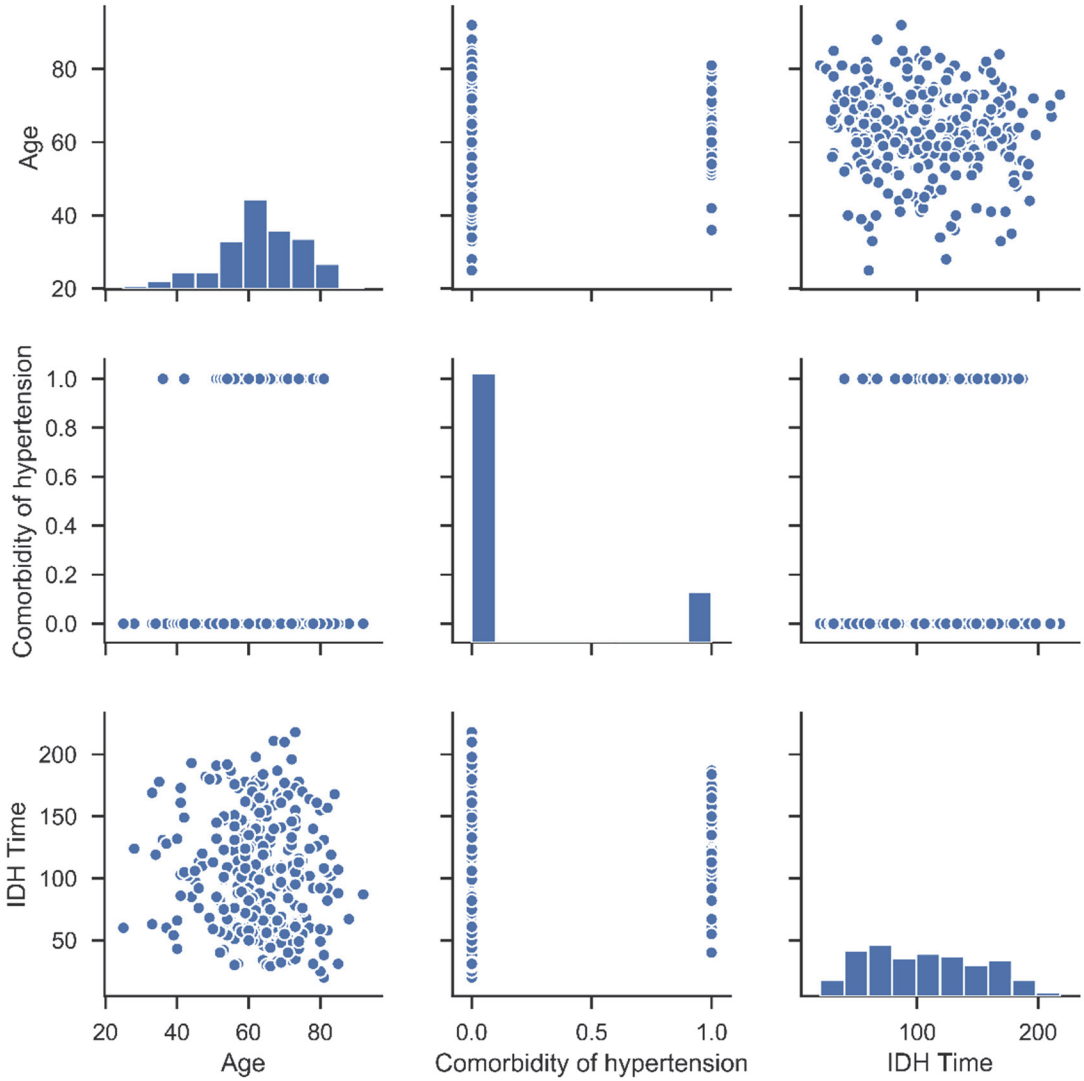


Figure 3-6. Scatterplot matrix for best model of 2-factor interaction with IDH time

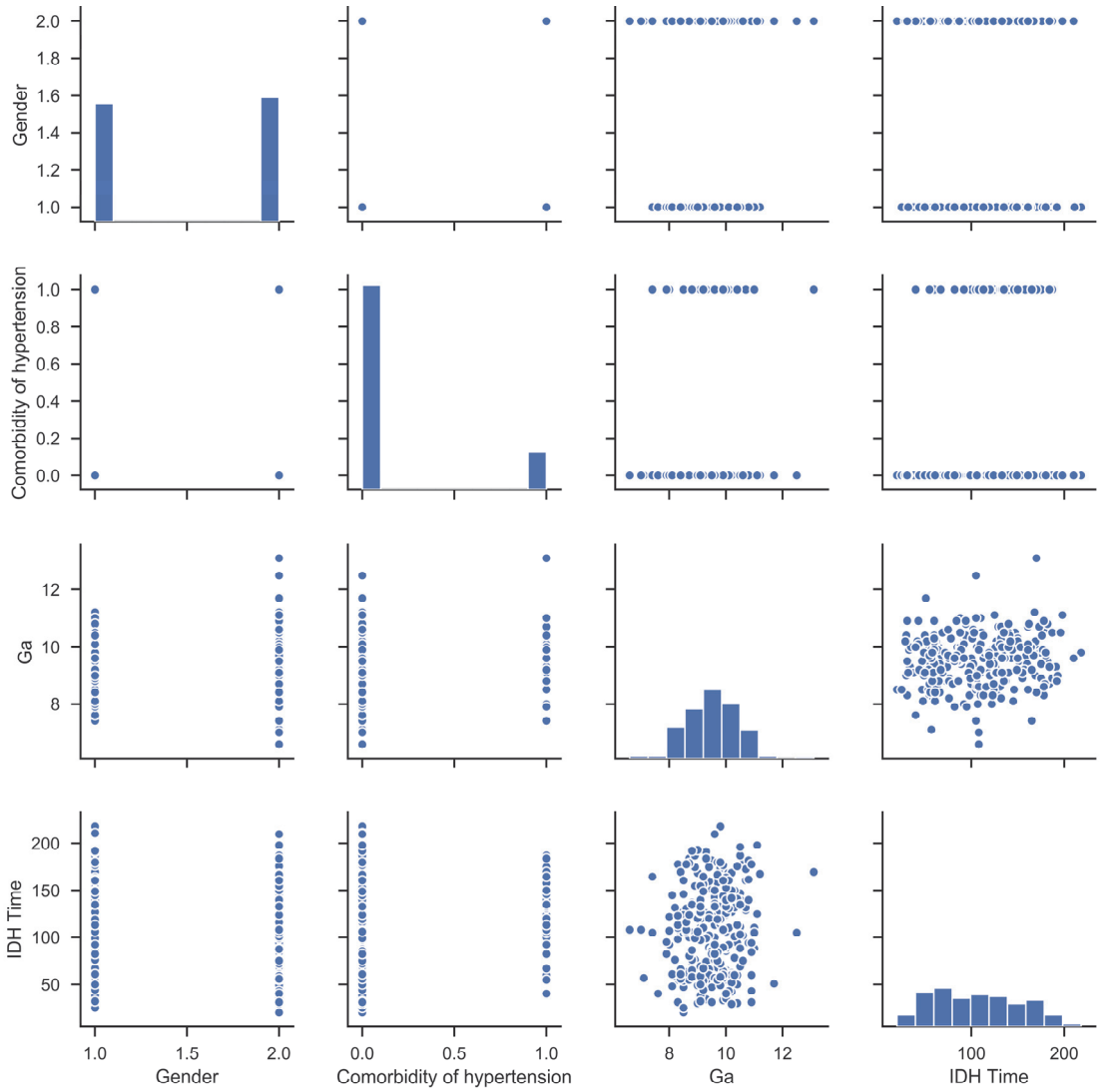


Figure 3-7. Scatterplot matrix for best model of 3-factor interaction with IDH time

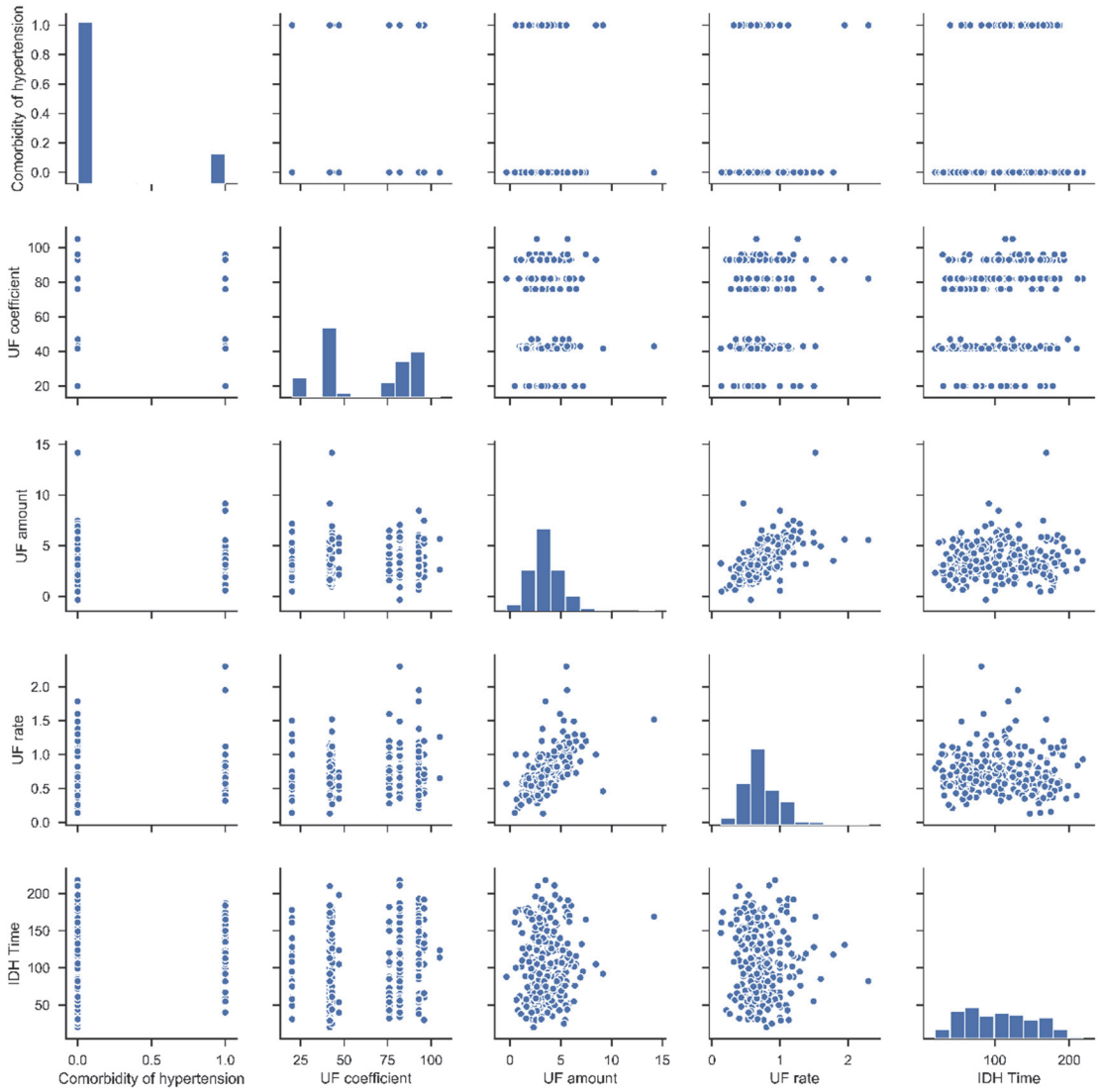


Figure 3-8. Scatterplot matrix for best model of 4-factor interaction with IDH time

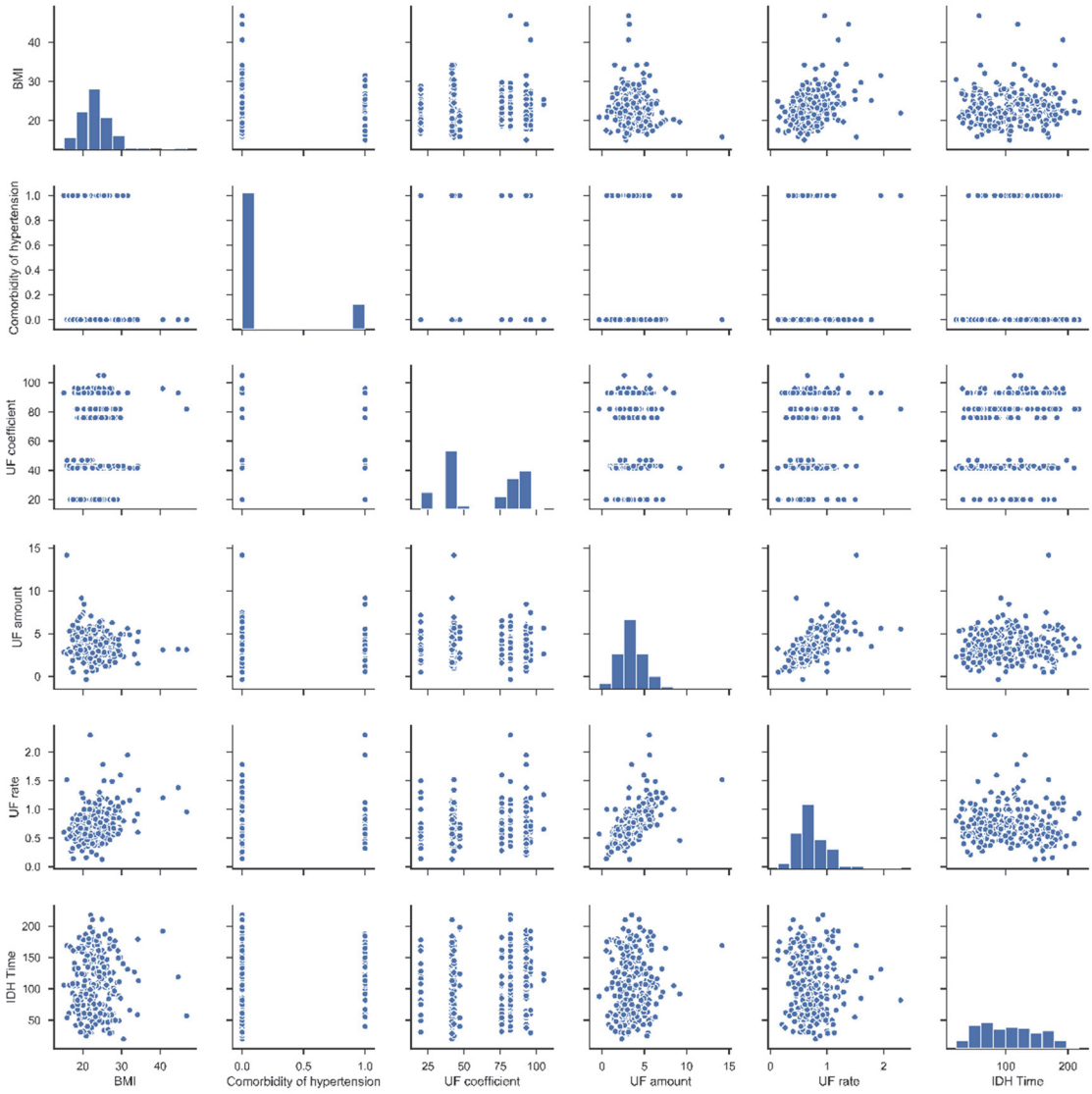


Figure 3-9. Scatterplot matrix for best model of 5-factor interaction with IDH time

Figure 3-10. Scatterplot matrix for best model of 6-factor interaction with IDH time

Figure 3-11. Scatterplot matrix for best model of 7-factor interaction with IDH time

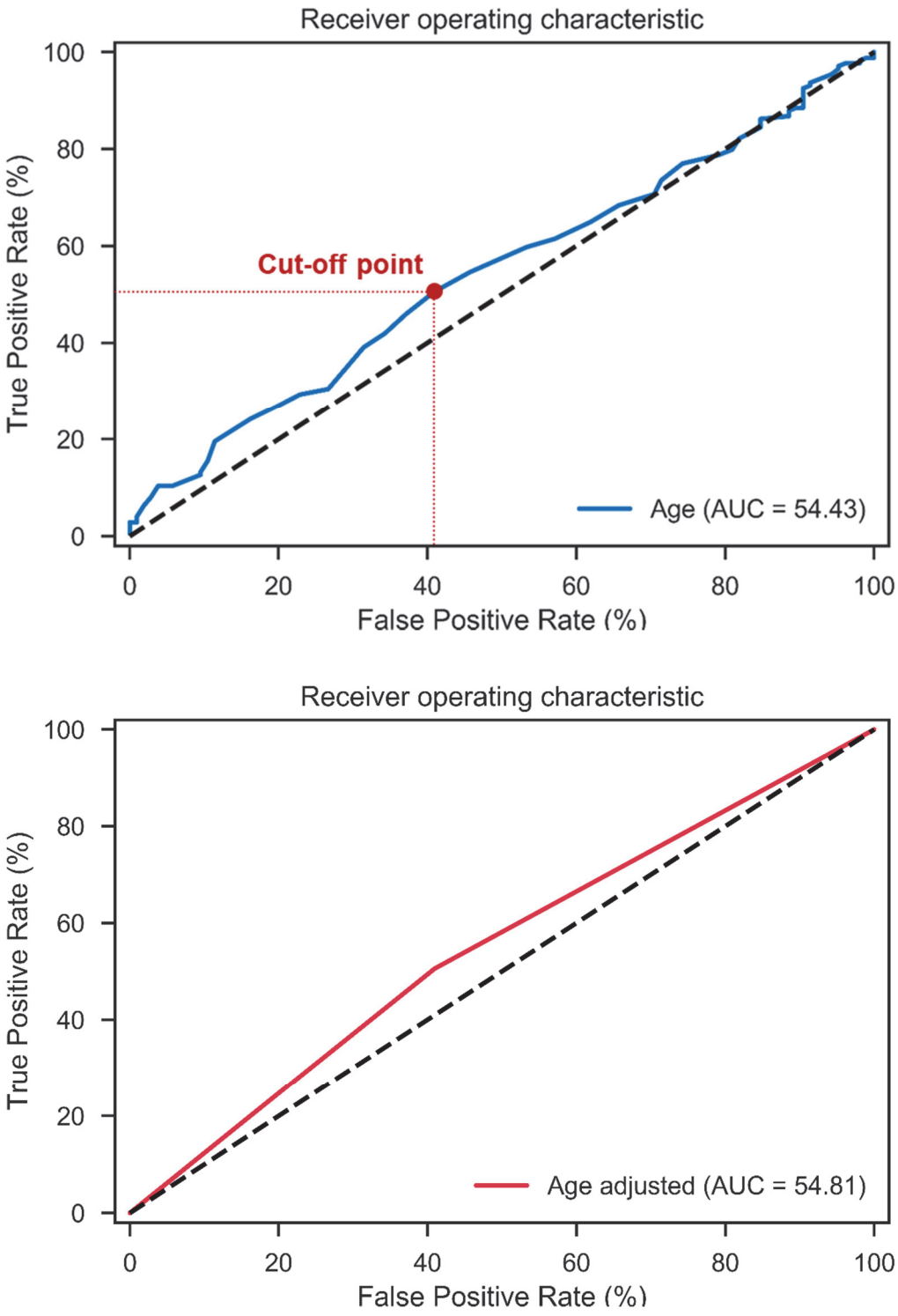


Figure 3-12. The best cutoff point of age according to ROC curve analysis

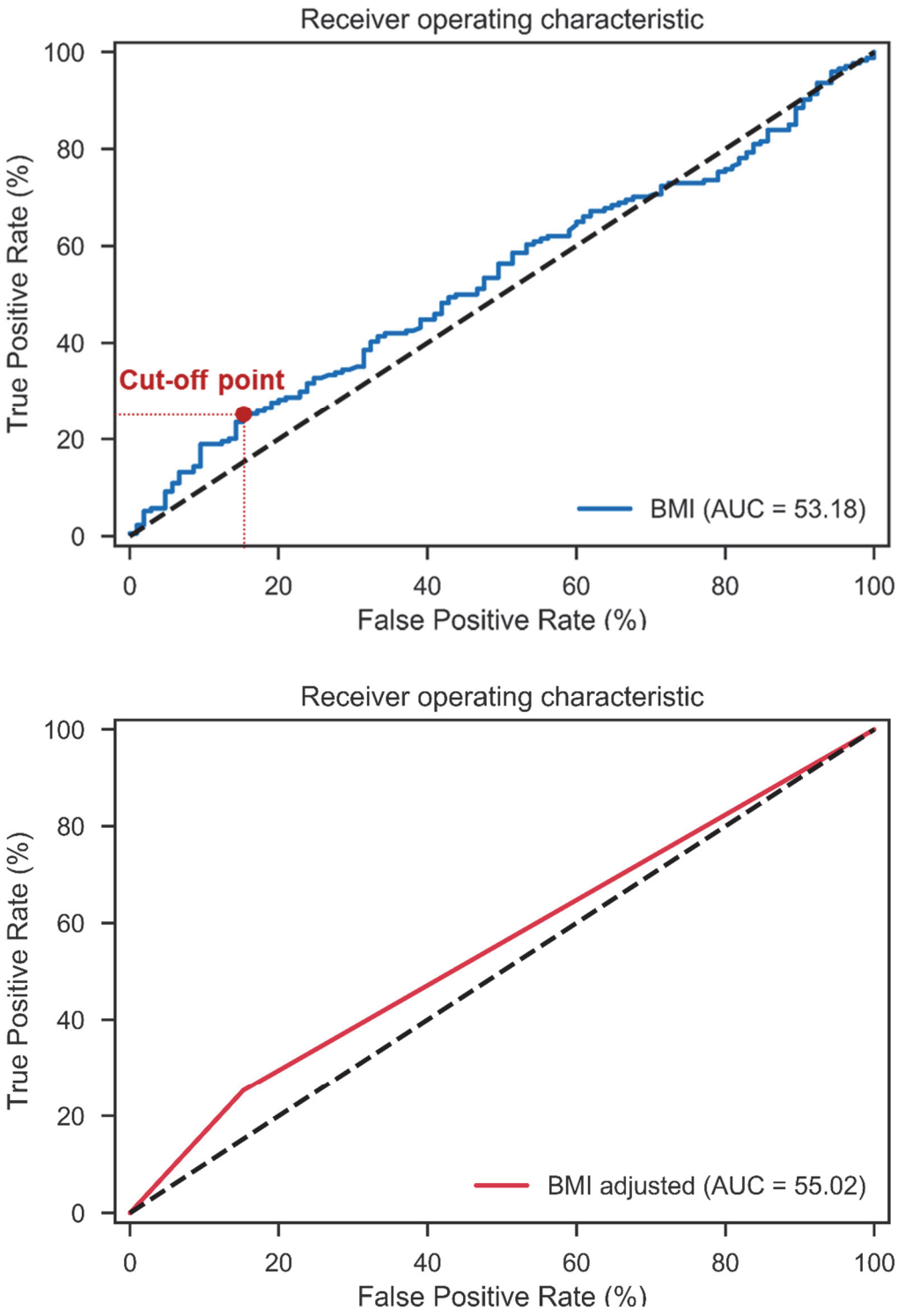


Figure 3-13. The best cutoff point of BMI according to ROC curve analysis

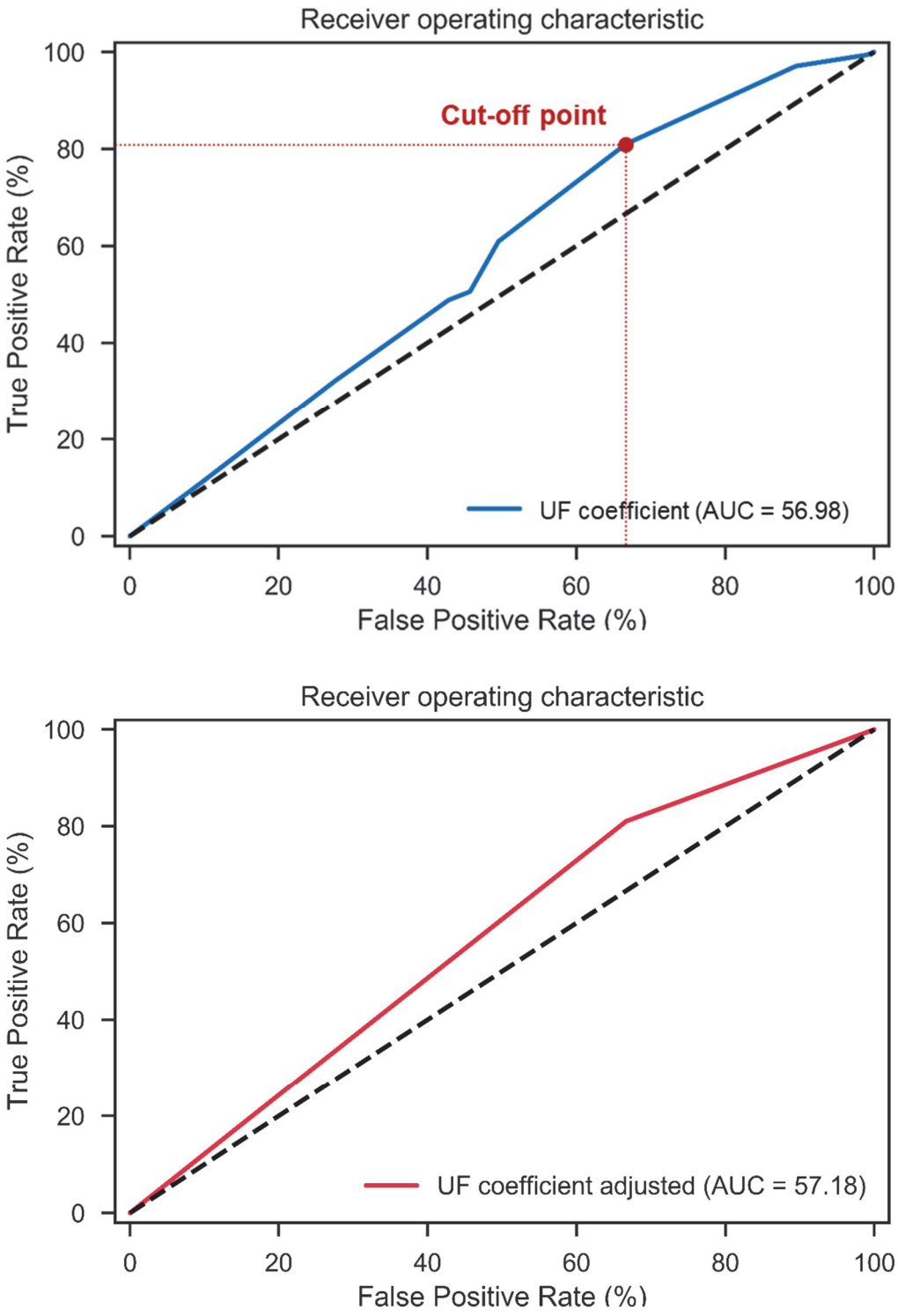


Figure 3-14. The best cutoff point of UF coefficient according to ROC curve analysis

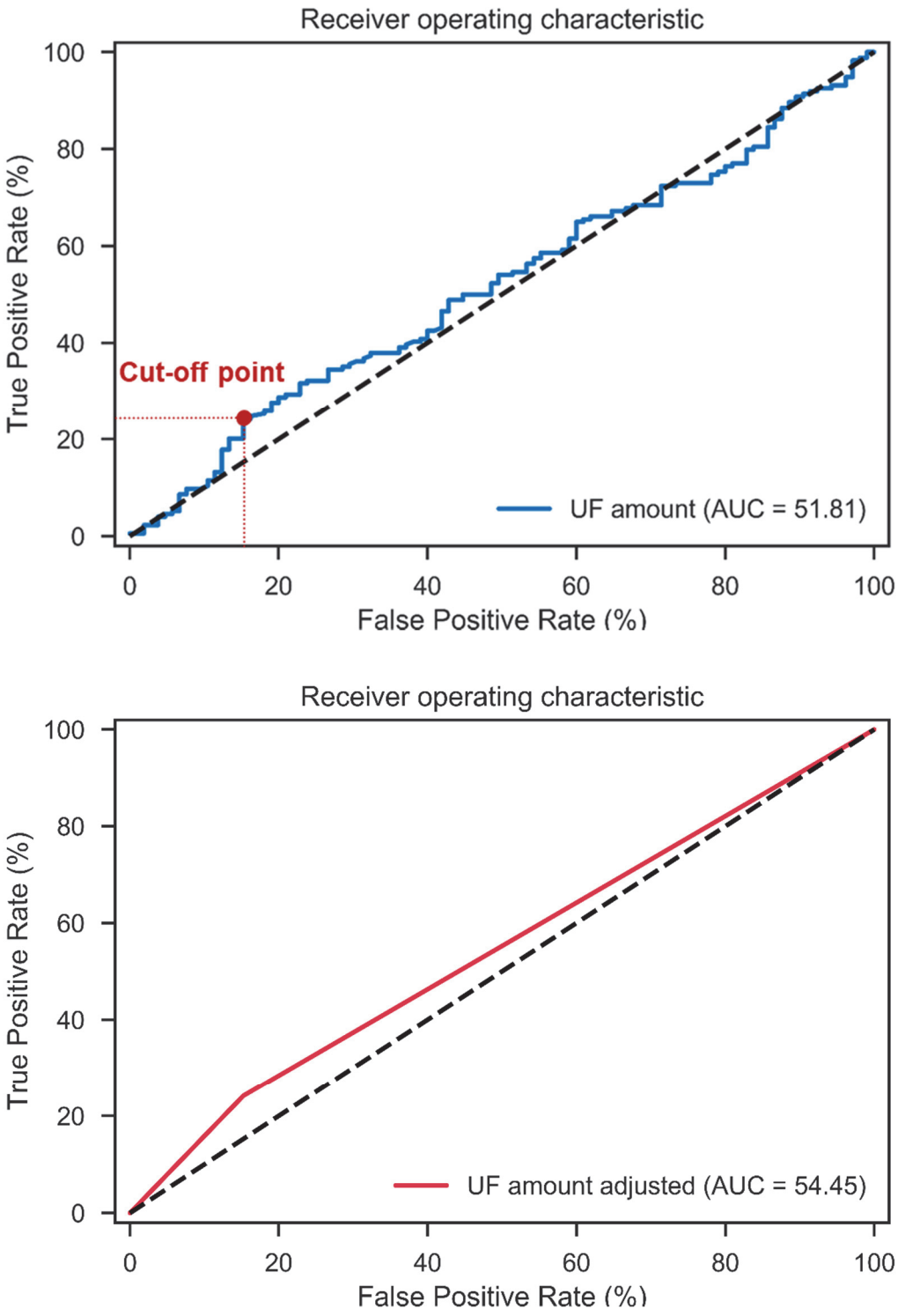


Figure 3-15. The best cutoff point of UF amount according to ROC curve analysis

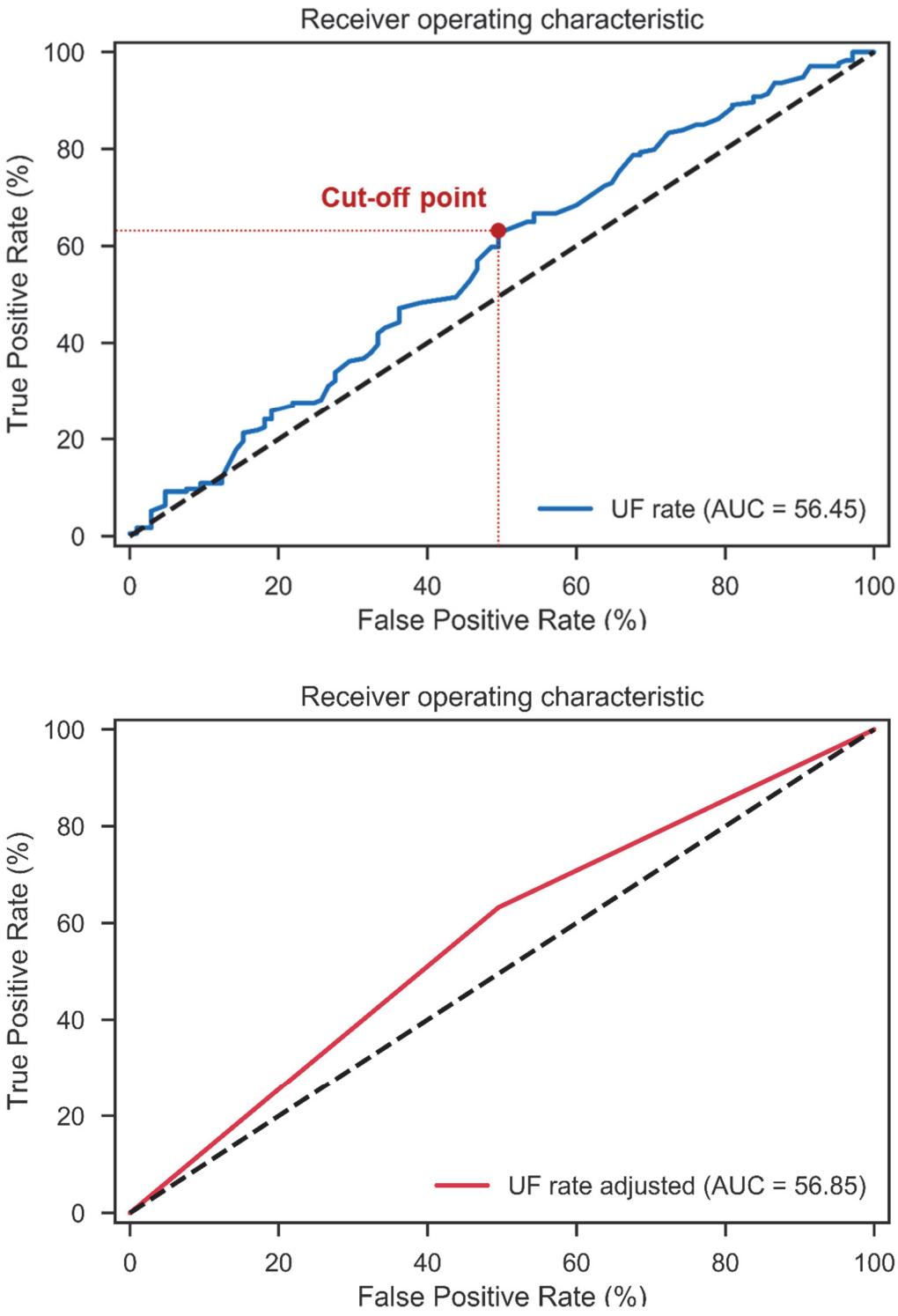


Figure 3-16. The best cutoff point of UF rate according to ROC curve analysis

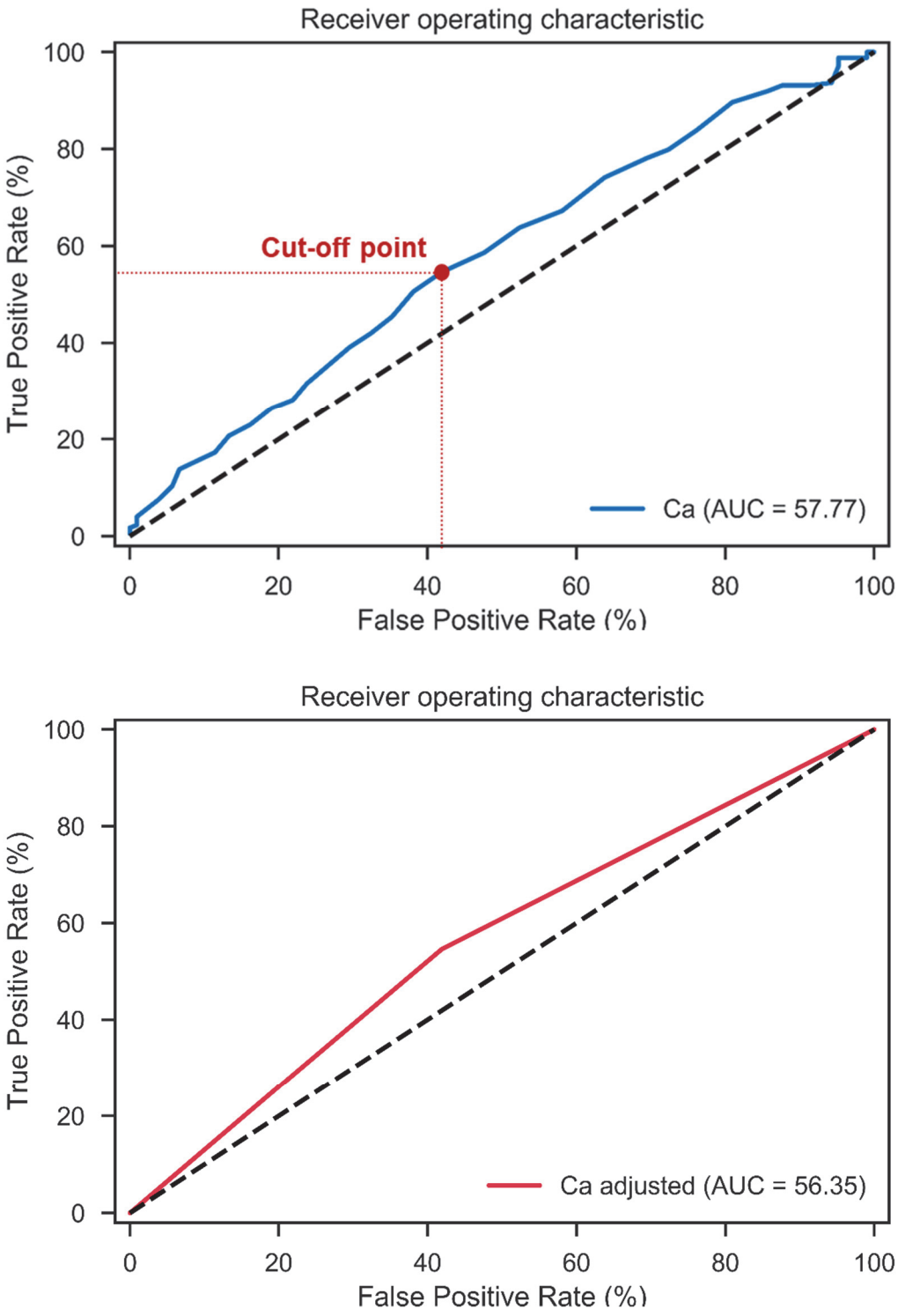


Figure 3-17. The best cutoff point of Ca according to ROC curve analysis

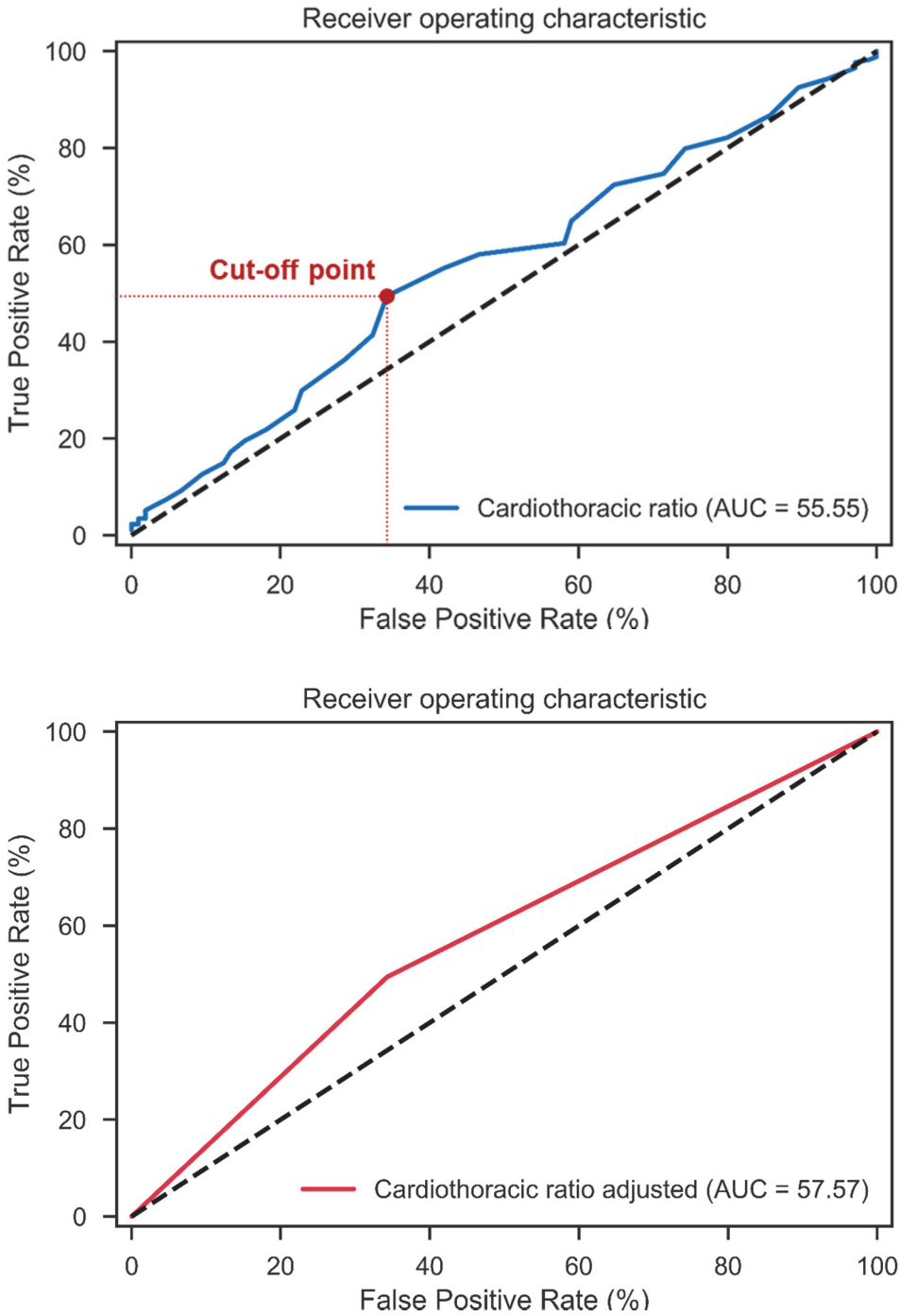


Figure 3-18. The best cutoff point of cardiothoracic ratio according to ROC curve analysis

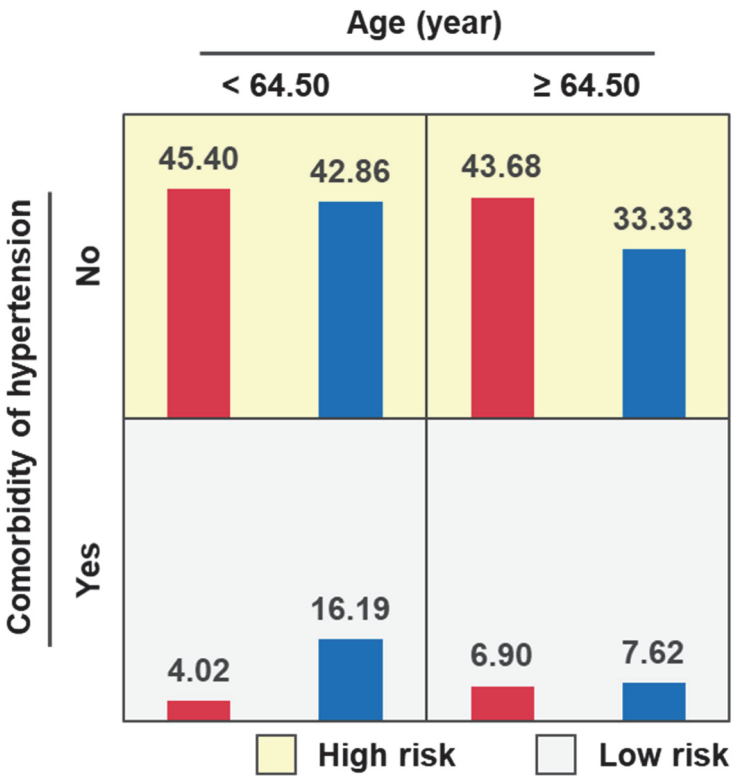


Figure 3-19. Summary of 2-factor combinations associated with high-low risk for IDH

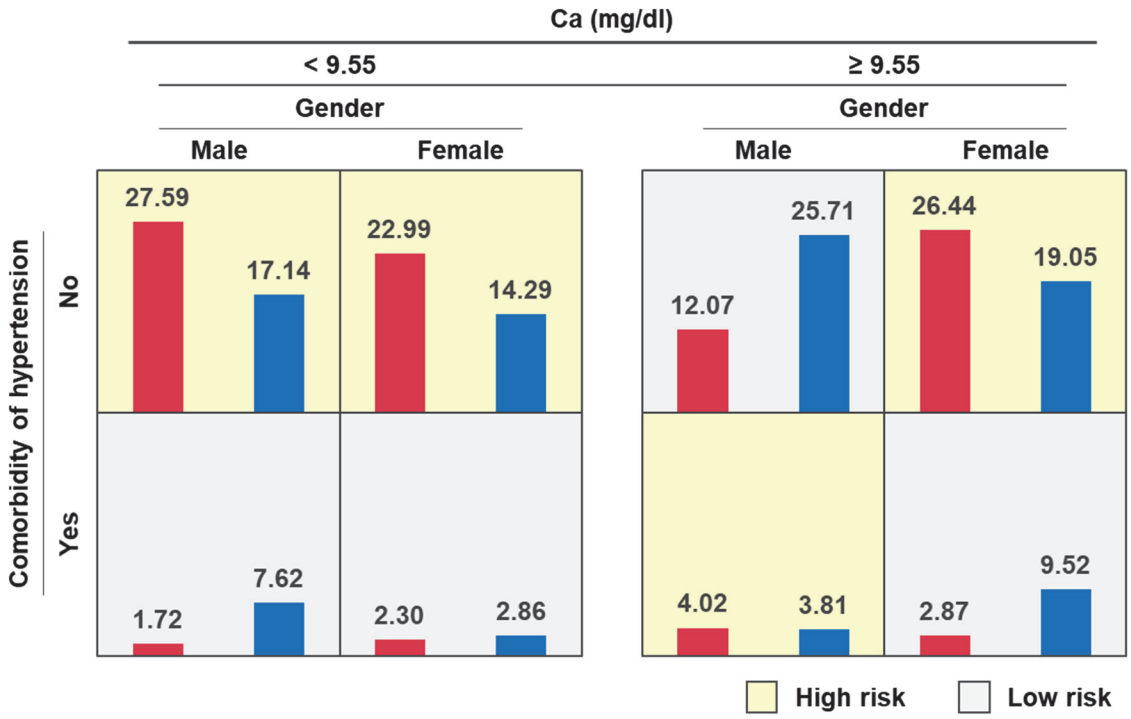


Figure 3-20. Summary of 3-factor combinations associated with high-low risk for IDH

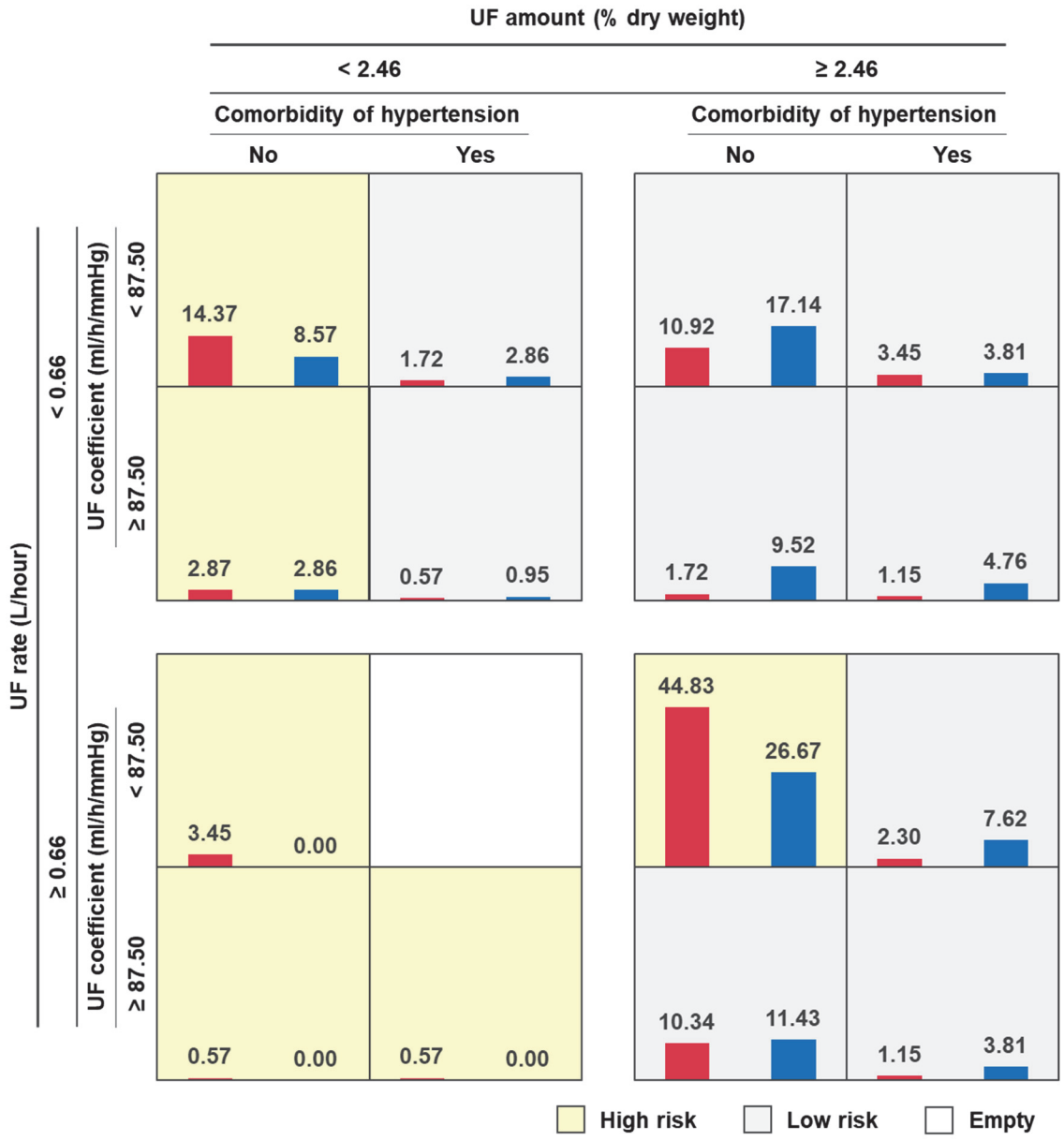


Figure 3-21. Summary of 4-factor combinations associated with high-low risk for IDH

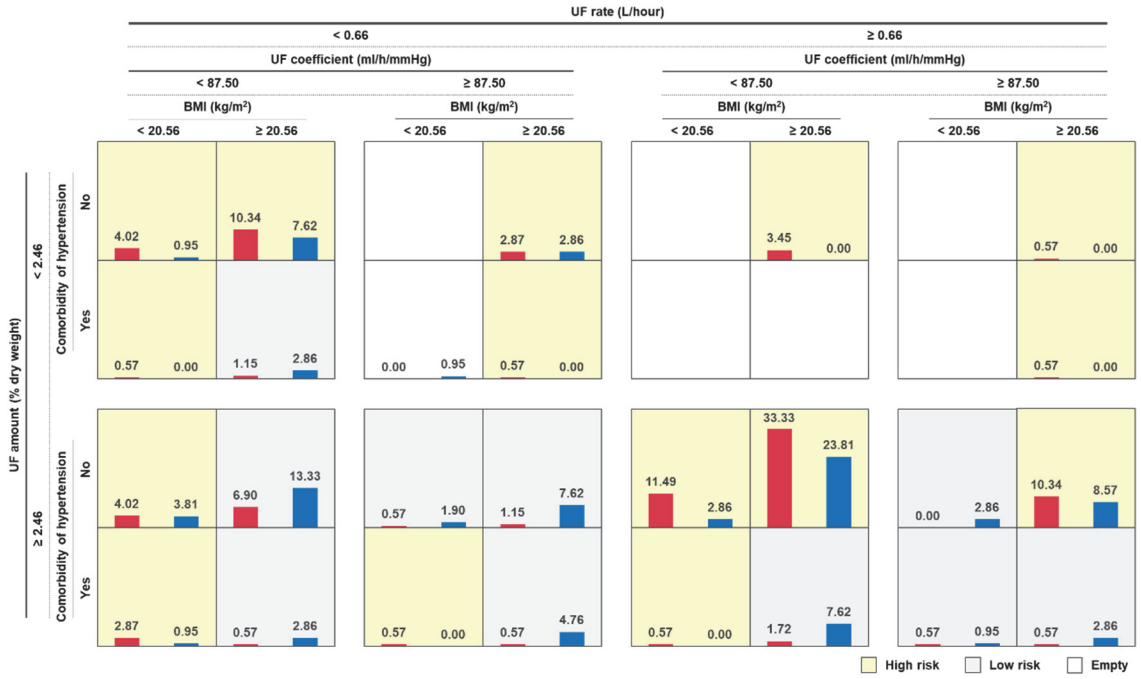


Figure 3-22. Summary of 5-factor combinations associated with high-low risk for IDH

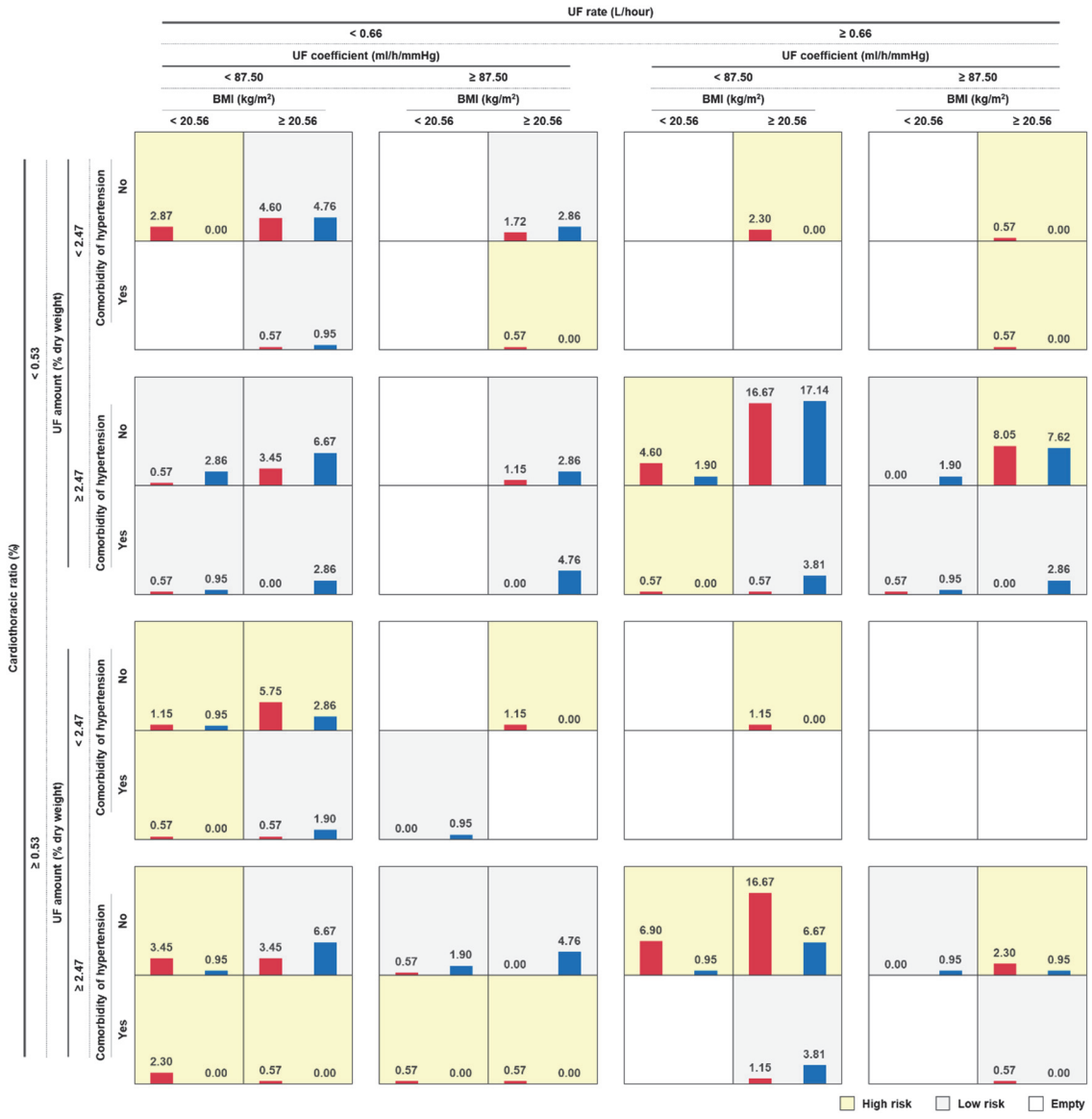


Figure 3-23. Summary of 6-factor combinations associated with high-low risk for IDH

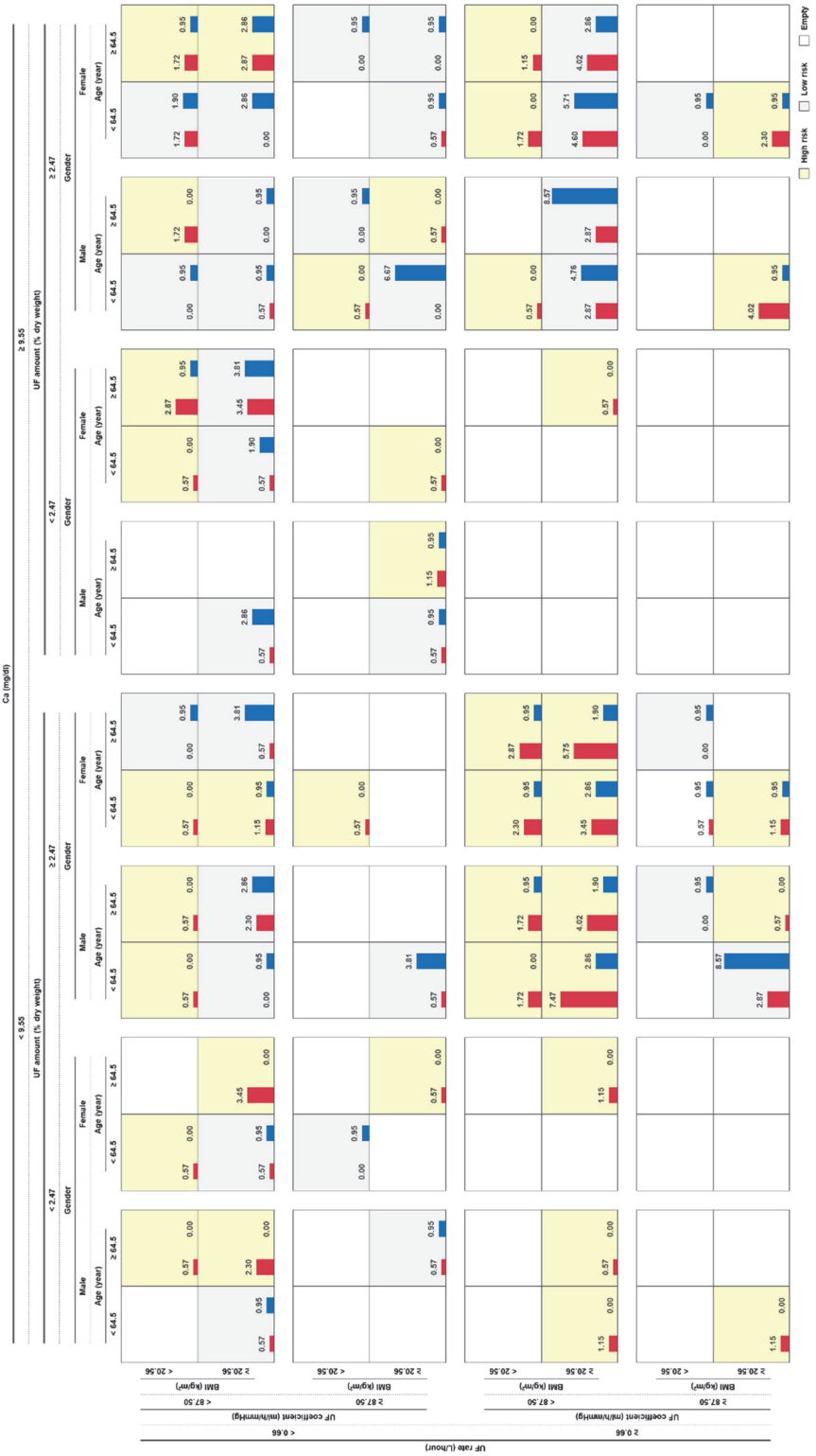


Figure 3-24. Summary of 7-factor combinations associated with high-low risk for IDH

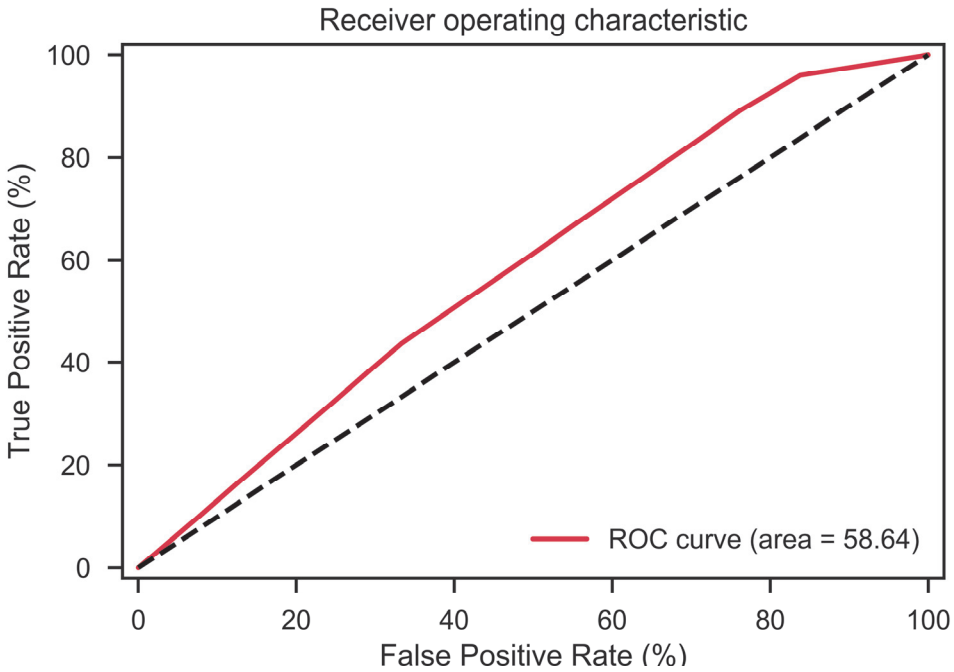


Figure 3-25. The 2-factor interaction to ROC curve analysis

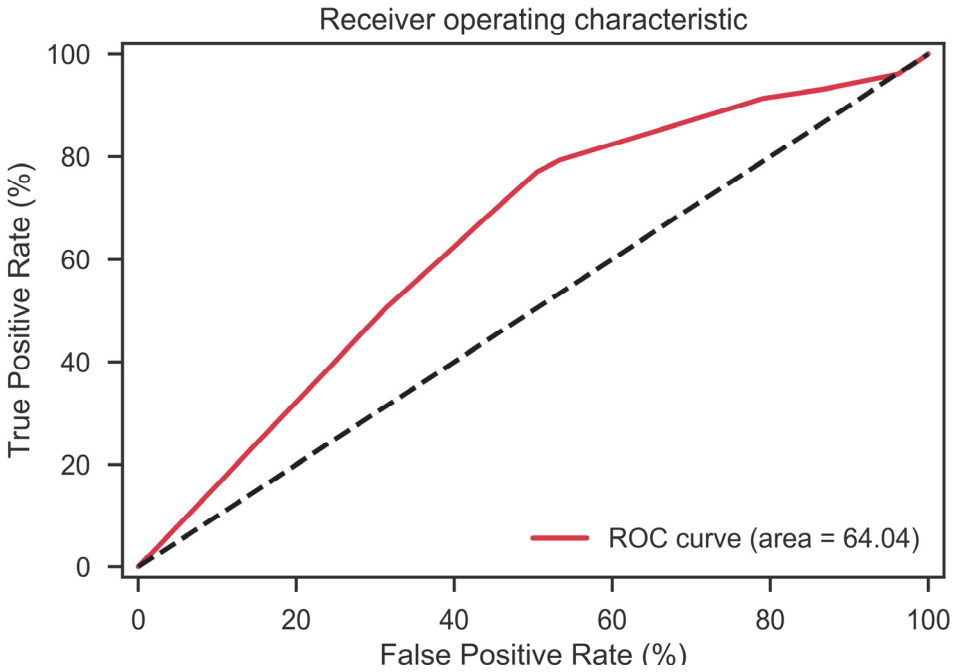


Figure 3-26. The 3-factor interaction to ROC curve analysis

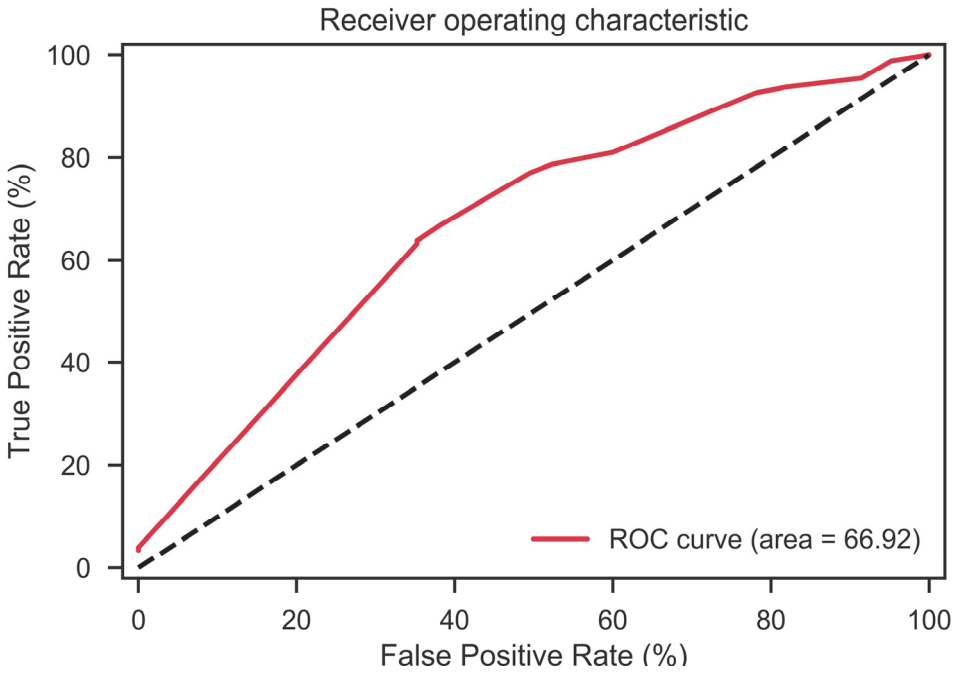


Figure 3-27. The 4-factor interaction to ROC curve analysis

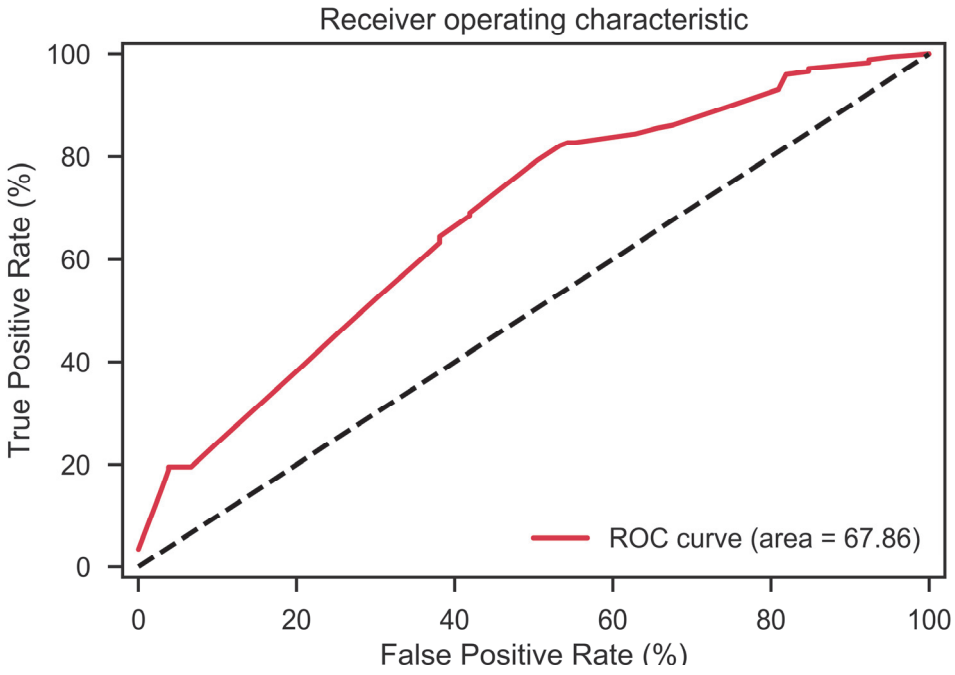


Figure 3-28. The 5-factor interaction to ROC curve analysis

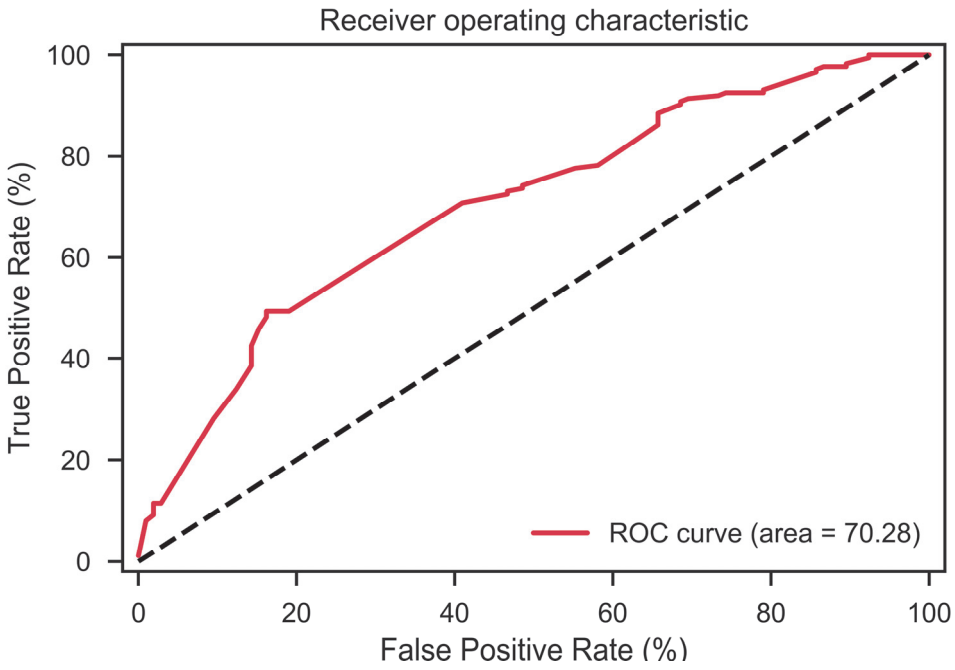


Figure 3-29. The 6-factor interaction to ROC curve analysis

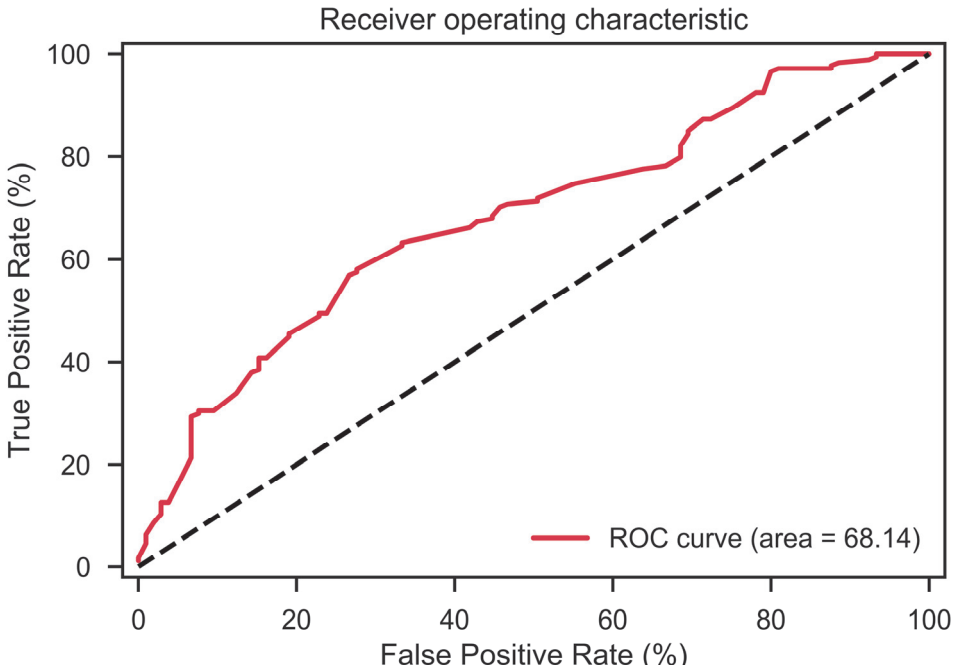


Figure 3-30. The 7-factor interaction to ROC curve analysis

3.4. Discussion

In the present study, we demonstrated that a deep learning can be used as a potential tool in clinical practice for the prediction of IDH during HD sessions. We examined the associations between several clinical variables and IDH by 7 models. The AUC was around 50 to 65 throughout these models. Although, the AUC did not reach a perfect standard, the clinical variables determined from the deep learning were reasonable, as per the experiences in the clinical practice. We found that IDH commonly occurred within the first 120 minutes of the HD initiation. We hypothesized that the intravascular refilling would not be adequate when the patients would undergo the UF procedure during the HD sessions. However, we could not determine the definitive cause of this time-effect of the occurrence of IDH because of incomplete evaluation of the individual cardiopulmonary function. Nevertheless, this time-effect of IDH cannot be refuted based on the common clinical experiences. We also found that leading factors associated with the occurrence of IDH during the HD were the UF amount (% dry weight), UF rate, and hypertension comorbidity. The full-adjusted model revealed positive correlations between the occurrence of IDH during the HD sessions and the BMI, UF amount, and hypertension

comorbidity. Accordingly, these results were compatible with the clinical experiences [65-69]. However, we cannot give the definitive reason for the correlation of the hypertension comorbidity. The supposed causes may include the use of antihypertensive during the pre-dialysis period, cardiac dysfunction related to the hypertension, etc. Furthermore, this also indicates the importance of being aware of the hypertension comorbidity during IDH management. The UF rate during the IDH event was inversely correlated with IDH. Rapid UF rate during the HD session is apt to elicit an inadequate intravascular refilling. Therefore, an optimal UF rate during the HD session should be anticipated to avoid the occurrence of IDH during the HD session. The full-adjusted model analyses revealed a positive correlation between the BMI levels and IDH during the HD sessions. However, their leading factors disappeared during the multi-factor interaction analyses. Therefore, we suggest that it is worthy to delineate their associations by performing large dataset analyses in the future.

This study considered 25 factors from the clinical data, which included the medical records and laboratory measurements. All the factors were tested and their multi-factor interaction were analyzed simultaneously using the DNN model. To our knowledge, the prediction of the occurrence of IDH as the outcome of the hemodialysis treatments has

not been previously used as a tuning factor in the computational models. The best model to analyze the multi-factor interaction was trained and validated by using the cross-validation approach. The multifactor interaction model achieved better accuracies than all-factor association model in short-term IDH occurrence. The 4-factor interaction model reported the highest performance in predicting the occurrence of IDH during hemodialysis. The 4-factor to 6-factor interaction models included the hypertension, UF rate, and UF amount as the factors for predicting IDH during hemodialysis, which is consistent with the full-adjusted multivariate regression analysis. Hence, the outcomes of the multi-factor interaction model may potentially contribute in the prediction of the occurrence of immediate IDH during hemodialysis.

There are limitations of using the deep learning in the present study. First, the included clinical variables could not cover the whole etiology of the occurrence of IDH during the HD sessions, such as, severe medical diseases that produce immediate hemodynamic changes when HD is initiated, time-effect of using unknown drugs in the patients, real time awareness of the occurrence IDH by the nursing staff, and other subtle conditions not identified by the deep learning. Second, the fundamental limitation arises from the nature of the deep networks, in which the neural network includes only the

clinical variables proposed by the medical staff. It is possible that several unknown variables might have been missed, and therefore, not included in the algorithm used in the present study. Meanwhile, it is noteworthy that understanding what kind of deep neural network to be used for the prediction of a clinical condition is still not an area of active research. We present the first study using a deep learning to predict the occurrence of IDH during the HD sessions. However, the accuracy rate could not achieve a satisfactory status. Hence, this algorithm cannot be used as a replacement to the comprehensive medical care during HD session. The validation of this algorithm requires further analyses involving a larger dataset, which may need a consensus from the experts on more confounding factors.

3.5. Summary

The present study demonstrated that a deep learning is a potential tool to determine the clinical factors associated with the occurrence of IDH during an HD session. The main goal of the future investigation may be to develop a satisfactory deep learning performance model based on the analyses of a larger dataset. To our knowledge, this is the first attempt of applying a DNN model including clinical variables to predict the occurrence of IDH during an HD session. In the future, we expect this model to achieve precision in predicting IDH by including more clinical samples and factors for analyses.

Chapter 4

DeepBarcode: DNA Barcodes Species Classification Using Deep Learning

DNA barcoding techniques were used for species identification with short fragment of sequences. Thanks to advances in sequencing technologies, DNA barcodes are emphasized gradually. The DNA sequences from different organisms are acquired easily and rapidly. Therefore, the DNA sequence analysis tools play an increasingly important role. This study presents DeepBarcode, a deep learning framework for species classification with DNA barcodes. DeepBarcode takes raw sequence data as input to transform one hot encoding as one dimension image and uses deep convolutional neural network with fully connected deep neural network for sequence analysis. It achieves over 90% average accuracies on simulation and real datasets. Although deep learning has obtained the outstanding performance for species classification with sequences, there are still significant challenges for its application. In conclusion, the DeepBarcode model is

an availably a potential tool for species classification, we expect that this model can contribute to understanding of DNA barcodes species identification.

4.1. Background

There are 8.7 million species on earth, approximately only 1.2 million species among them were compiled fully by taxonomic classification. Recently, loss of biodiversity has been emphasized as a major global environmental problem, ecologists are continuously devoted to reforming strategies for biological conservation and natural resources protection. However, the main obstruction to access taxonomic classification of biosphere is often attributed to the problem of taxonomic impediment. In this problem, taxonomic information are not always accessible and comprehensible for researchers who are not taxonomic experts. Therefore, the genetic information and specifically DNA sequences was proposed to overcome this problem [70, 71].

DNA barcoding technique was first proposed for species identification in 2003, which used mitochondrial cytochrome c oxidase subunit I (COI) as DNA tags to identify biological specimens in the animal kingdom. A shorter DNA barcode sequence provides adequate messages to distinguish specimen into species, this technique has been demonstrated that it was successfully and applicably applied for species discrimination and identification. With the growth of COI barcodes in animal group, other group were

emphasized gradually, such as chloroplast ribulose-bisphosphate carboxylase gene (*rbcL*) and maturase K (*matK*) for plant and internal transcribed spacer (*ITS*) for fungus. The fundamental issue of DNA barcoding is that, once robust reference sequence database is created, it will easily obtain sufficient information and assign a query sequence to classify correct species from unknown specimens. Consequently, several approaches for species identification with DNA barcode were proposed, it can be categorized into three methods, namely tree-based taxonomic methods (e.g. Neighbor Joining [72]), similarity-based taxonomic methods (e.g. BLAST [73]), character-based taxonomic methods (e.g. BLOG [74]) and machine learning-based taxonomic methods (e.g. supervised learning classification [75, 76]).

The identification species through analysis of DNA sequences from organism is a big challenge in bioinformatics. Many machine learning approaches for identification species with DNA barcode were proposed, such as probabilistic methods, unsupervised clustering and supervised classification [74-76]. In DNA barcode classification problem, a *reference* library as known species is obtained from DNA barcode specimen. A *query* sets as unknown species is collected by DNA barcode sequences. Conversion of *reference* set and *query* set into supervised learning are given training set and testing set respectively.

The training set consists of specimens with priori known species with labels, and the testing set consists of specimens which is unknown species for classification [74, 75]. In supervised learning classification approach, the training set is used to train a suitable model for classification then testing set is used to test the performance of trained model, consequently, the classification accuracy is obtain. Several classifiers are suggested for species classification with DNA barcode, such as support vector machine (SVM), k-nearest neighbor (KNN), decision tree (DT), naïve Bayes (NB), multi-layer perceptron (MLP), and random forest (RF) [74, 75].

In machine learning field, the state-of-the-art deep learning has been actively studied and has been achieved record-breaking performance in various applications such as bioinformatics problems (involve image processing [52], signal processing [54] and sequence analysis [77]). One of deep learning model is called convolutional neural network (CNN) which was demonstrated that can achieve superior performance in various prediction task, images and sequences processing especially. CNN can adapt to the recognition task which learns from the data during training procedure, rather than select hand-crafted feature extraction/selection or preprocessing recipes by experts. CNNs capture both local and global textures as important features from training data via

its advantage of the characteristics of local connecting, parameters sharing, pooling and multi-layers using [43].

This paper introduces a CNN model that automatically captures the textures of DNA sequences from training data to predict the species through the performance comparison with the *ad-hoc* method of supervised learning. The *ad-hoc* supervised learning methods were implemented in this study, including SVM, KNN, NC, NB, DT, MLP and RF classifiers in Scikit-Learn library. We collected simulation data and real DNA barcode sequences for species identification, the real world DNA barcode sequences involved COI, *rbcL*, and *ITS* genes. In species identification task, our proposed CNN architectures took superior accuracy and exceeded the classification performance of others supervised learning approaches, which represents that our proposed method to provide a useful architectures in DNA barcode sequences analysis.

4.2. Materials and methods

4.2.1. Dataset

Simulation data

The Coalescent package was used to generate simulation DNA barcode sequences in Mesquite software, the datasets were available at http://dmb.iasi.cnr.it/supbarcodes.php. Three datasets were simulated according to the Yule model [75] which generated 50 species by random ultrametric species tree, and considered two terms: time of species divergence and effective population size (N_e). 20 individuals per species were generated with gene tress using N_e = 1,000, 10,000 and 50,000 that replicated 100-fold scheme resulted in 300 datasets in total. The parameter N_e made the complexity of dataset increasingly. The sequence length of 650 base pair (bp) were setting similar to standard DNA barcode size in real world. **Table 4-1** summarizes simulation data information.

Real data

Six real datasets were obtained from GenBank Nucleotide Database, organized source data are available at http://dmb.iasi.cnr.it/supbarcodes.php. Those data possessed three properties including high phylogenetic diversity, lack of inter-specific sequence differences (it leads identification complexity) and different genomic compartments [75]. The information of six selected datasets are shown in **Table 4-2** including *Cypraeidae*, *Drosophila*, *Inga*, *Bats*, *Fishes*, and *Birds* with DNA barcode sequences.

Table 4-1. Summary of the simulated dataset

Dataset	N_e	**Individual***	**Seq. length***	**Species***
Ne1000	1,000	20	650	50
Ne10000	10,000	20	650	50
Ne50000	50,000	20	650	50

Legend: Individual* indicates number of sequences for each species; Seq. length* is the length of the sequences; Species* represents number of species (i.e. classes).

Table 4-2. Real dataset summary

Dataset	Num. of sequences*	Length*	Species*	Gene region*
Cypraeidae	1656 / 352	614	211	COI
Drosophila	499 / 116	663	19	COI
Inga	786 / 122	1,838	63	tmTD, ITS
Bats	695 / 144	659	96	COI
Fishes	515 / 111	718	82	COI
Birds	1306 / 317	691	150	COI

Legend: Num. of sequences* represents the number of sequences that divided into training set and testing set; Length* is the length of the sequences; Species* indicates the number of species (i.e. classes); Gene region* indicates that DNA barcodes were obtained from the gene region(s).

4.2.2. Convolutional neural network architecture

The deep convolutional neural network (CNN) is a well-known deep learning architecture that was first proposed by LeCun *et al*. [21, 44] in 1989. The basic components of CNN consist of input, convolution, pooling, fully-connected and output, layer by layer. Given n training data with labels of m classes for input that represents as $\{\boldsymbol{X}_{(n)}, y_{(n)}\}$, where $\boldsymbol{X}_{(n)}$ are matrices of size n get $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$ with $y = \{y_1, y_2, \ldots, y_n\}$, $y_i \in \{1, 2, \ldots, m\}$.

Convolutional layer

The convolutional layer is used as filters to play a role of weight sharing and feature extracting. The l-th convolutional operator computes as:

$$\boldsymbol{z_C} = f_{conv}(\boldsymbol{X}) \tag{6}$$

where,

$$f_{conv}(\boldsymbol{X})_{lk} = \mathrm{ReLU}\left(\sum_{i}^{M-1}\sum_{j}^{N-1} K_{i,j}^{k} X_{l+i,j} + b_l\right) \tag{7}$$

where K denotes a $M \times N$ matrix of convolutional kernel, k is the index of kernels. M is the window size and N is the number of input channels. b_l denotes the bias. The ReLU

(rectified linear unit) is a non-linearity function that sets all negative values to zero, the formula is shown as follow:

$$\text{ReLU(x)} \begin{cases} x \; if \; x \geq 0 \\ 0 \; if \; x < 0 \end{cases} \tag{8}$$

Pooling layer

The max-pooling layer is used to reduce the size of the output from prior layer (i.e. convolutional layer). The max-pooling operation simply select the maximum value in a window of neurons cluster at output of convolutional layer, the *l*-th max-pooling operator computes as:

$$\boldsymbol{z_P} = f_{maxpool}(\boldsymbol{z_C}) \tag{9}$$

where,

$$f_{maxpool}(\boldsymbol{z_C})_{lk} = \max\left(\left\{\boldsymbol{z}_{ClM,k}, \boldsymbol{z}_{C(lM+1),k}, \dots, \boldsymbol{z}_{C(lM+M-1),k}\right\}\right) \tag{10}$$

where *k* is the index of kernels. *M* is the window size.

Fully connected layer

Next, prior tuned parameters fed into the fully connected layer (i.e. deep neural network, DNN) from convolutional layer and pooling layer. The DNN considers three

layers including input layer, hidden layer and output layer for classification is an objective function shown as follow:

$$y = \text{argmax}\, f(\boldsymbol{w}x + \boldsymbol{b}) \tag{11}$$

where $\boldsymbol{w}$ and $\mathbf{b}$ are weight and bias matrix in the model. x is a testing vector data.

The prior output of feature map are flattened into the input layer of DNN. In the hidden layer, h neurons are considered for computing a non-linear transformation using ReLU activation function that computes as follows:

$$h_i = \max(0, \boldsymbol{w}_i x + \boldsymbol{b}_i) \tag{12}$$

where i is i-th neuron, $\boldsymbol{w}_i$ and $\boldsymbol{b}_i$ denote weight and bias vector of hidden neuron h_i, respectively. The h neurons of hidden layer as output matrices map to next i-th neuron of output layer is represented as follows:

$$o_i = f(\boldsymbol{w}_i h + \boldsymbol{b}_i) \tag{13}$$

where $\boldsymbol{w}_i$ and $\boldsymbol{b}_i$ denote weight and bias vector of neuron o_i in output layer, and the function $f(.)$ is the *softmax* function that computes the probabilities of all possible cases for output interpretation. The probability is calculated as follow:

$$p(y|x) = \frac{1}{\sum_{j=1}^{m} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_m^T x^{(i)}} \end{bmatrix} \tag{14}$$

where $\theta_1, \theta_2, \ldots, \theta_m$ are parameters (i.e. weight and bias) learned by automatic model, $x^{(i)}$ denoted *i*-th input. The best parameter of neural network model are optimized by minimizing cross entropy loss function on the entire training data $\{\boldsymbol{X}_{(n)}, y_{(n)}\}$, defined as:

$$\mathrm{L}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} loss\big(f\big(X_{(i)}\big), y_{(i)}\big), \tag{15}$$

The adam optimizer algorithm are used for loss function optima finding through back-propagation step. The update formula of parameter θ are described as:

$$\theta \leftarrow \theta - \eta \frac{\partial \mathrm{L}(\theta)}{\partial \theta} \tag{16}$$

where η is the learning rate.

4.2.3. DNA barcodes species classification

Figure 4-1 illustrates the progress of five steps in DeepBarcode architecture for species classification with DNA barcode sequences. Step 1) data processing: the DNA barcode sequences can be downloaded from GenBank. Those sequences are aligned with sequence alignment tool using Muscle in MEGA [78] software, because of the different length of sequence from different association number in GenBank. Step 2) the DNA sequences include four elements A, C, G, T that are encoded to one hot encoding as a

special one dimensional image for next step. Step 3) the convolutional layer and max-pooling layer are constructed to compute prior sequences of one hot encoding. In this step, n filters of $m \times m$ kernel size and $l \times l$ windows of pooling size are used that results k matrices with ReLU activation function and then k matrices is flattened into a vector for next step. Especially, the convolutional layer and max-pooling layer can set to multi-layer layer by layer. Step 4) the fully connected deep neural network is implemented to integrate prior features and to enhance the feature extraction in subsampling. Here, the adam optimizer is used for finding loss function optimization, and the ReLU activation function is used again for each neuron. Step 5) lastly, *softmax* function is used as classifier to compute the probability of species as output and then each species is correctly classified. The DeepBarcode is implemented with Keras toolkit (available on https://keras.io/) which is based on Tensorflow [63] libraries in Python language.

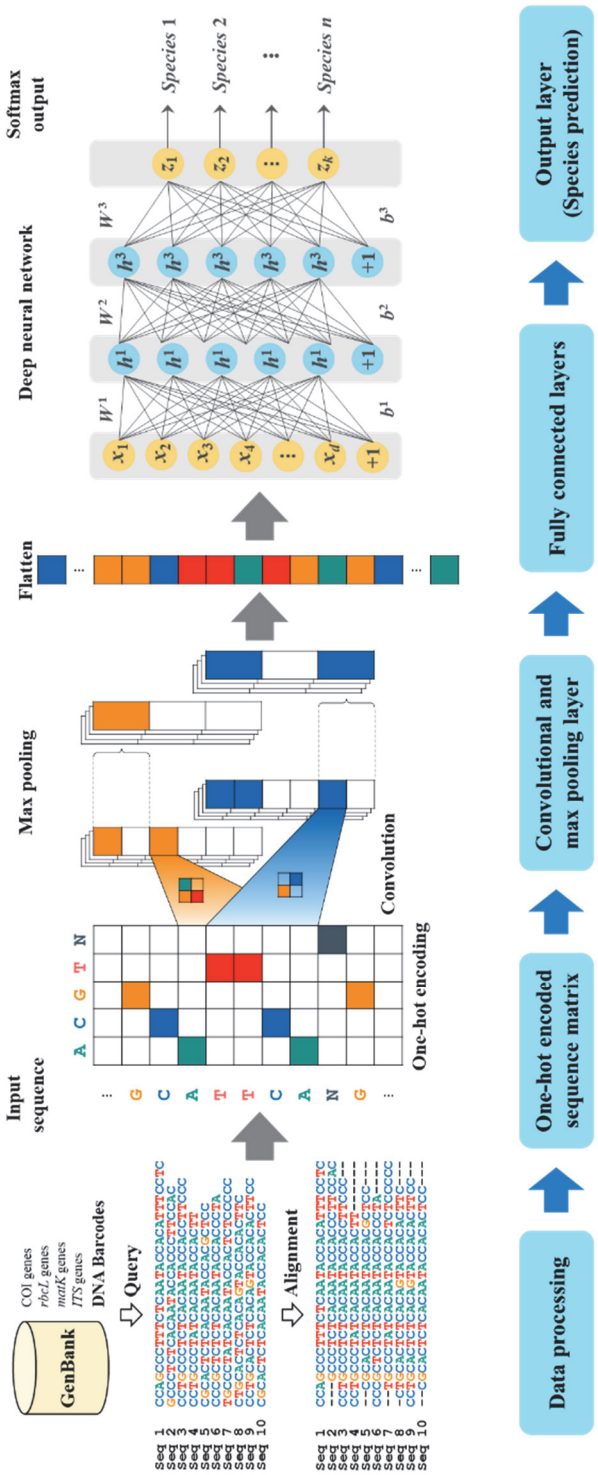


Figure 4-1. An illustration of a deep neural network with three hidden layers

4.3. Results

4.3.1. Experiment of performance comparison on simulation data

Table 4-3 shows a comparison of classification accuracies obtained by classifiers taken from DeepBarcode, RF, SVM, KNN, DT, NB and MLP methods. All the classifiers were applied to three simulated data with DNA barcode sequences and independently executed 100 times (from 100-fold scheme) for each dataset. In Table 3, the average accuracy is 95.84%±1.37%, 95.44% ± 1.38%, 95.34% ± 1.66%, 94.44% ± 1.60%, 94.05% ± 1.69%, 93.39% ± 1.75% and 91.52% ± 2.50% for the DeepBarcode, RF, SVM, KNN, DT, NB and MLP method in simulated data, respectively. Each classifier has over 90% accuracy that indicates this data distribution fitting those classifier on their hypothesis. Consequently, DeepBarcode had highest performances on those three simulation data.

Table 4-3. Performance comparison on empirical datasets (%)

Method	Dataset			
	Ne1000	Ne10000	Ne50000	Average
DeepBarcode	**96.74±1.67**	**96.57±1.16**	**94.21±1.27**	**95.84±1.37**
RF	96.65±1.64	96.42±1.03	93.25±1.48	95.44±1.38
SVM	96.33±1.92	96.09±1.49	93.60±1.58	95.34±1.66
KNN	96.66±1.70	95.75±1.39	90.91±1.71	94.44±1.60
DT	95.99±1.65	94.93±1.55	91.24±1.88	94.05±1.69
NB	96.31±1.86	95.70±1.26	88.15±2.13	93.39±1.75
MLP	94.72±2.65	92.95±2.34	86.89±2.50	91.52±2.50

Legend: DeepBarcode, proposed method; RF, random forest; SVM, support vector machine; KNN, k-nearest neighbor; DT, decision tree; NB, naïve Bayes; MLP, multi-layer perceptron. The best performances on each dataset are in bold.

4.3.2. Experiment of performance comparison on real data

Six real datasets with DNA barcode sequence were tested for classification task. The real data were divided into training data and testing data by 80%-20% split respectively. In **Table 4-4**, the experimental results show the average classification accuracy is 97.61% ± 2.24% for DeepBarcode and that is better than others. DeepBarcode, RF and SVM achieved 100% classification accuracy on Fishes datasets that revealed the training data completely fitting for testing data on trained models. However, it is worth to discussing data collection. Perhaps, we can collect further data on fish species in the further. It's worth noting that the NB and KNN classifiers obtained adverse results on Birds dataset, it indicates not all of supervised learning classifiers are suitable for DNA barcode classification.

Table 4-4. Performance comparison on real datasets (%)

Dataset	MLP	NB	DT	KNN	SVM	RF	DeepBarcode
Cypraeidae	94.89	91.48	93.75	95.17	95.17	96.02	**96.31**
Drosophila	**99.14**	**99.14**	97.41	**99.14**	**99.14**	**99.14**	**99.14**
Inga	81.97	73.77	86.89	89.34	**93.44**	92.62	**93.44**
Bats	**99.31**	97.92	95.83	95.14	**99.31**	98.61	**99.31**
Fishes	98.20	97.30	96.40	97.30	**100.00**	**100.00**	**100.00**
Birds	90.54	53.63	91.80	69.72	94.64	90.54	**97.48**
Average	94.01	85.54	93.68	90.97	96.95	96.16	**97.61**
SD	6.19	16.68	3.55	9.97	2.60	3.51	**2.24**

Legend: DeepBarcode, proposed method; RF, random forest; SVM, support vector machine; KNN, k-nearest neighbor; DT, decision tree; NB, naïve Bayes; MLP, multi-layer perceptron. The best performances on each dataset are in bold.

4.4. Discussion

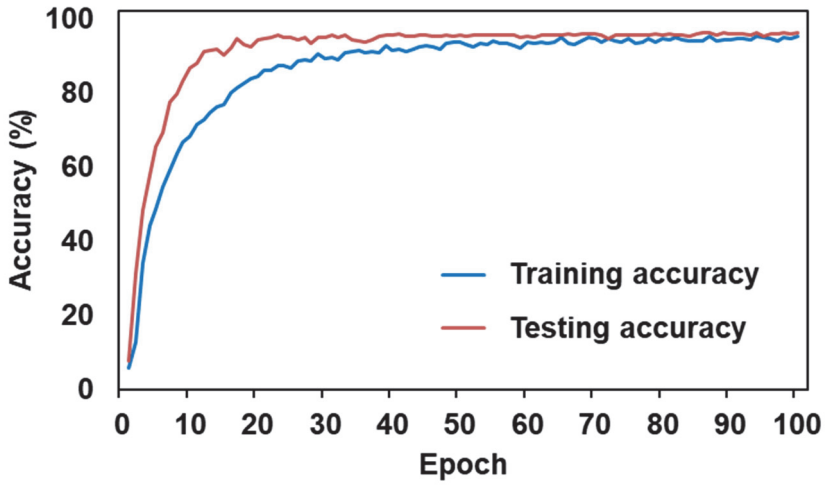
The CNN architecture was used in this study, four characteristics of CNN include local connected, weight sharing, convolution and pooling, and multi-layer [43]. The CNN is a non-linear multi-layer transformation approach that learns important feature automatically from data instead of manual feature extraction from experts in the field. Many researches have been demonstrated that CNN achieves state-of-the-art performance in image classification and recognition [79, 80]. Particularly, the DNA sequences can be represented as an image to analysis by CNN model. A variety of sequence analysis issues have demonstrated the CNN model achieves excellent performance on sequence analysis, such as non-coding RNA sequences [81], non-coding sequence variants on DNA methylation [77], DNA- and RNA-binding proteins from next generation sequencing technique [82], protein sequences [83] and genome sequences [84-87]. Consequently, this study used the CNN model in DeepBarcode for species classification with DNA barcode, and the better performances than other classifiers on experimental datasets.

The overfitting problem is a frequent concern with large complexity of computational model. The deep learning techniques are notoriously difficult to train,

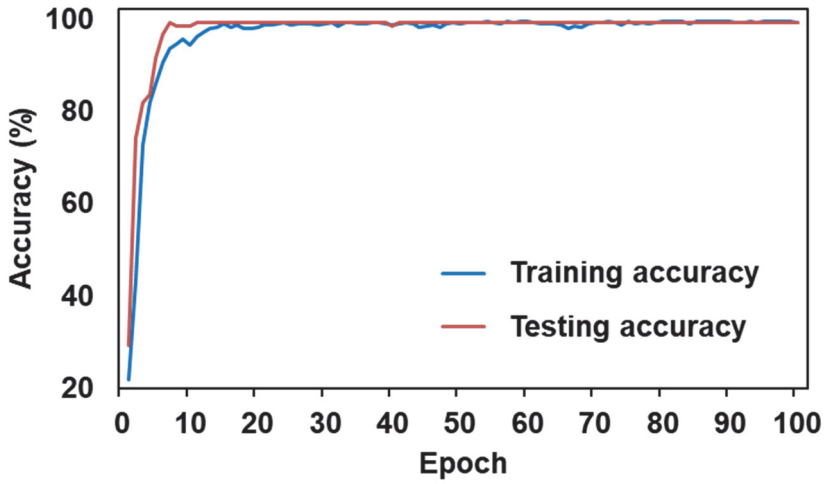
several new regularization strategies were proposed in order to train a robust and reliable model, such as dropout, early stopping, batch normalization, cross-validation and data augmentation [88]. Dropout technique was used in this study which is a simple and common regularization way to avoid overfitting. Dropout randomly removed units from the model during training progress that can decrease the model complexity and improve generalization [40]. **Figure 4-2** shows the overfitting of observation on six real data. It reveals the DeepBarcode without overfitting problem.

The main goal of DNA barcoding is to be an effective tool for species identification that uses the short DNA sequence (e.g. COI, *rebL* gene) designing a cheap, intuitive and straightforward barcode. Those reliable short DNA barcodes reduced storages on the database and improved species of economic [72, 89]. In addition, the single-nucleotide polymorphism barcodes (SNP tags) were proposed for barcoding which more easily obtained, quickly stored and continuously reduced storage cost. Resulting, the DNA barcoding could speed up the discovery of new or unknown species [90, 91]. However, the precision and rapid identification and classification for taxonomy on DNA barcode is still a large challenge. In this study, the simulation and real data were tested for species classification on DeepBarcode model that obtained the outstanding performance. With

the advantages of DNA barcoding, we expect the DeepBarcode as a high potential tool for DNA barcoding issues in the future.

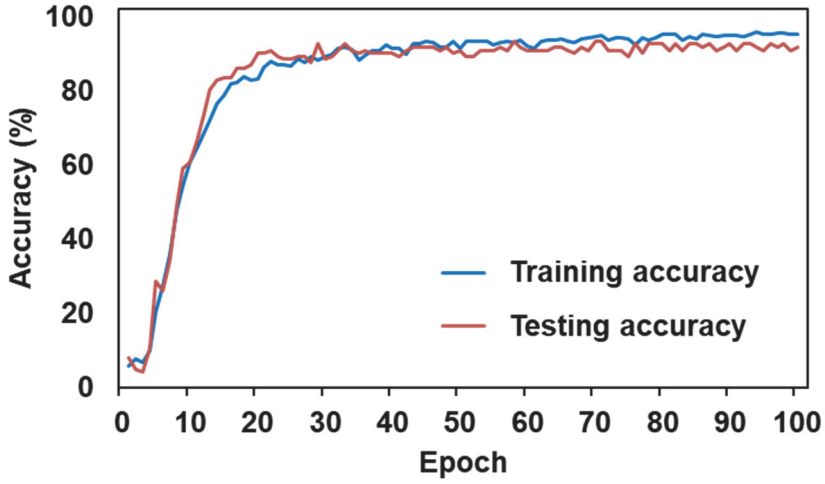


(a) Cypraeidae dataset

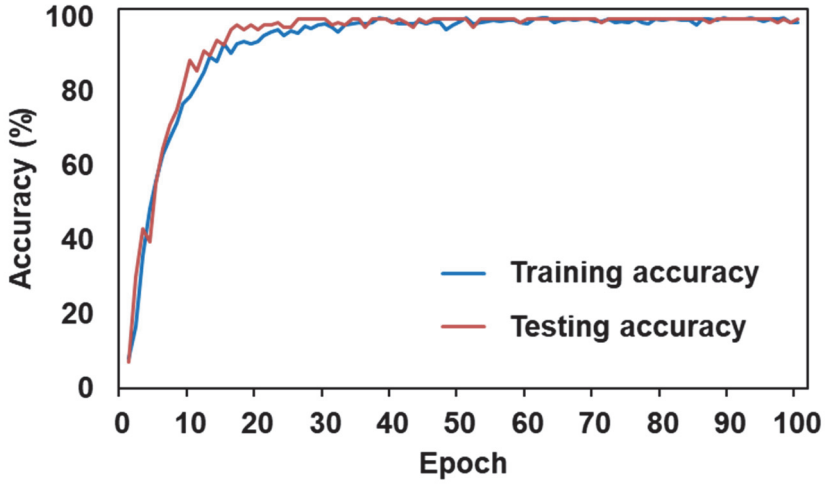


(b) Drosophila dataset

Figure 4-2. A comparison of the training set and testing set accuracy on real datasets

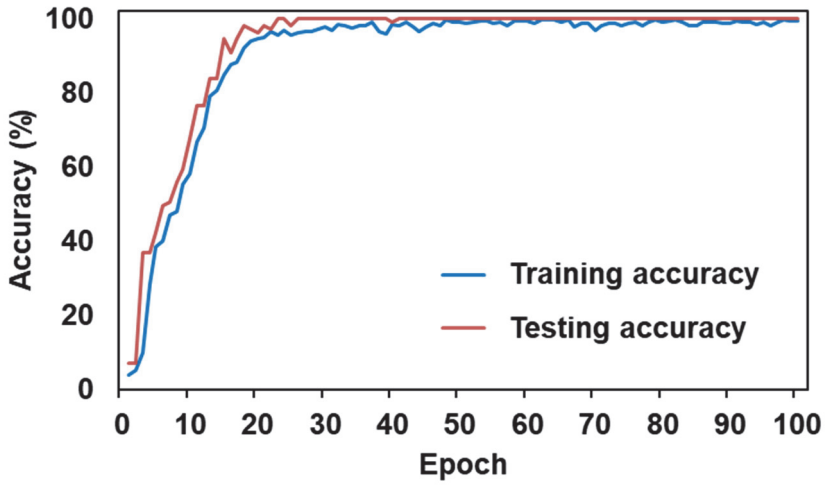


(c) Inga dataset

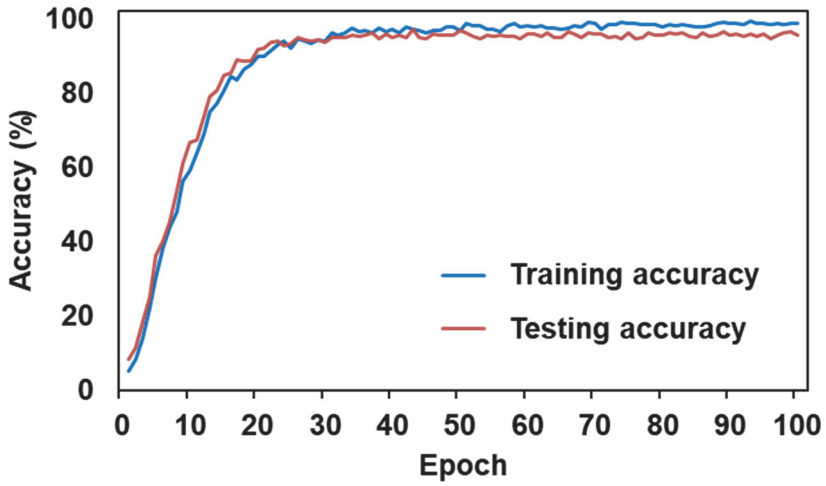


(d) Bats

Figure 4-2. *continued*



(e) Fishes dataset



(f) Birds dataset

Figure 4-2. *continued*

4.5. Summary

The DeepBarcode is based on supervised learning. Therefore, our model relies on the well-known specie sequences, availability, and quality of training set that limits to expand new species. However, this paper provides exhaustive methods for species classification that names an unknown specimen to a known species via trained model with its DNA barcodes. Although deep learning has improved the pleasurable performance for species classification, there are still significant challenges for its application in DNA barcodes analysis. On the whole, techniques like DeepBarcode maintain the potential to lead researchers to advance investigation of DNA barcoding, which can conduce understanding of DNA barcodes species identification gradually.

Chapter 5

Conclusions and Future Work

The conventional machine learning has been successfully applied to biomedical informatics, it is anticipated that this application will create an exciting new trend in deep learning. Expectably, two topics of deep learning techniques were implemented and discussed in this thesis and it had outstanding performances on two experiments of intradialytic hypotension occurrence prediction and species classification in this thesis. At present, although deep learning had completed many ad hoc applications and provided excellent results, that remains several potential challenges, such as suitable model selection, hyperparameter, domain knowledge exposition of deep learning results and limited or imbalanced data. Besides, as deep learning pleasurable performance enhances, complex computation increases, acceleration of deep learning needs supplementary studies.

In the future work, firstly, the intradialytic hypotension occurrence prediction still deserve further study, including more collections of clinical character and samples, and immediate analysis the patients' situation in hemodialysis. Also, one or more active and novel models will be considered natural to develop for more different domains in biomedical informatics in future research. Secondly, the sequence analysis using deep learning had rosy performance on the species classification with DNA barcodes. The barcode of life data system had collected an ocean of DNA barcodes over 6 million barcodes with toward approximately two hundred thousand animal species, sixty thousand plant species and twenty thousand other species. Therefore, it is still an important issue to discuss various species classification, more convenience tools or applications in DNA barcoding. In addition, the next generation sequencing technique is popular and hot field, sequence analysis approaches may be use in that study. Finally, many models and hyperparameters of deep learning techniques need to be tuned for various problems, the optimization algorithm like genetic algorithm and particle swarm optimization algorithm can be used to model and parameter optimization.

References

[1] C. A. Kulikowski, E. H. Shortliffe, L. M. Currie, P. L. Elkin, L. E. Hunter, T. R. Johnson*, et al.*, "AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline," *J Am Med Inform Assoc,* vol. 19, pp. 931-938, 2012.

[2] E. V. Bernstam, J. W. Smith, and T. R. Johnson, "What is biomedical informatics?," *J Biomed Inform,* vol. 43, pp. 104-110, 2010.

[3] W. Hersh, "A stimulus to define informatics and health information technology," *BMC Med Inform Decis Mak,* vol. 9, Art. no. 24, 2009.

[4] E. H. Shortliffe and J. J. Cimino, *Biomedical informatics: computer applications in health care and biomedicine*: Springer Science & Business Media, 2013.

[5] M. A. Musen and J. H. van Bemmel, *Handbook of medical informatics*: Bohn Stafleu Van Loghum Houten, the Netherlands, 1997.

[6] R. A. Greenes and E. H. Shortliffe, "Medical informatics. An emerging academic discipline and institutional priority," *JAMA,* vol. 263, pp. 1114-1120, 1990.

[7] J. H. Van Bemmel, "The structure of medical informatics: Bibliography on educational courses at the Free University, Amsterdam," *Med Inform,* vol. 9, pp. 175-180, 1984.

[8] Y. Peng, Y. Zhang, and L. Wang, "Artificial intelligence in biomedical engineering and informatics: an introduction and review," *Artif Intell Med,* vol. 48, pp. 71-73, 2010.

[9] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychol Rev,* vol. 65, pp. 386-408, 1958.

[10] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans Inf Theory,* vol. 13, pp. 21-27, 1967.

[11] G. V. Kass, "An exploratory technique for investigating large quantities of

categorical data," *J R Stat Soc Ser C-Appl Stat,* vol. 29, pp. 119-127, 1980.

[12] L. Breiman, *Classification and regression trees*. New York: Routledge, 1984.

[13] P. E. Utgoff, "Incremental induction of decision trees," *Mach Learn,* vol. 4, pp. 161-186, 1989.

[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature,* vol. 323, pp. 533-536, 1986.

[15] R. E. Schapire, "The strength of weak learnability," *Mach Learn,* vol. 5, pp. 197-227, 1990.

[16] J. R. Quinlan, *C4. 5: programs for machine learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[17] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans Neural Netw,* vol. 5, pp. 157-166, 1994.

[18] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J Comput Syst Sci,* vol. 55, pp. 119-139, 1997.

[19] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, Montréal, Qué, Canada, 1995, pp. 338-345.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Mach Learn,* vol. 20, pp. 273-297, 1995.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc IEEE,* vol. 86, pp. 2278-2324, 1998.

[22] L. Breiman, "Random forests," *Mach Learn,* vol. 45, pp. 5-32, 2001.

[23] G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput,* vol. 18, pp. 1527-1554, 2006.

[24] R. Raina, A. Madhavan, and A. Y. Ng, "Large-scale deep unsupervised learning using graphics processors," *Proceedings of the 26th Annual International*

Conference on Machine Learning, Montreal, Quebec, Canada, 2009, pp. 873-880.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[26] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche*, et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature,* vol. 529, pp. 484-489, 2016.

[27] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez*, et al.*, "Mastering the game of Go without human knowledge," *Nature,* vol. 550, pp. 354-359, 2017.

[28] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, Berkeley, Calif., 1967, pp. 281-297.

[29] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika,* vol. 32, pp. 241-254, 1967.

[30] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J Cybernetic,* vol. 3, pp. 32-57, 1973.

[31] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J R Stat Soc Ser B-Stat* vol. 39, pp. 1-38, 1977.

[32] W. H. E. Day and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *J Classif,* vol. 1, pp. 7-24, 1984.

[33] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans Speech Audio Process,* vol. 3, pp. 72-83, 1995.

[34] C. Yizong, "Mean shift, mode seeking, and clustering," *IEEE Trans Pattern Anal Mach,* vol. 17, pp. 790-799, 1995.

[35] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, 1996, pp. 226-231.

[36] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases," *Sigmod Rec,* vol. 25, pp. 103-114, 1996.

[37] B. Scholkopf, A. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput,* vol. 10, pp. 1299-1319, 1998.

[38] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science,* vol. 315, pp. 972-976, 2007.

[39] C. Angermueller, T. Parnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Mol Syst Biol,* vol. 12, p. 878, 2016.

[40] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo*, et al.*, "Deep learning for health informatics," *IEEE J Biomed Health Inform,* vol. 21, pp. 4-21, 2017.

[41] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Brief Bioinform,* vol. 18, pp. 851-869, 2017.

[42] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Brief Bioinform,* Art. no. bbx044, 2017.

[43] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, pp. 436-444, 2015.

[44] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard*, et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput,* vol. 1, pp. 541-551, 1989.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput,* vol. 9, pp. 1735-1780, 1997.

[46] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. Cambridge, MA, USA: MIT press, 2016.

[47] Y. Bengio, "Learning deep architectures for AI," *Foundat and Trends Mach Learn,* vol. 2, pp. 1-127, 2009.

[48] R. Agarwal, "How can we prevent intradialytic hypotension?," *Curr Opin Nephrol Hypertens,* vol. 21, pp. 593-599, 2012.

[49] R. F. Reilly, "Attending rounds: a patient with intradialytic hypotension," *Clin J Am Soc Nephrol,* vol. 9, pp. 798-803, 2014.

[50] Q. Zhu, X. Li, A. Conesa, and C. Pereira, "GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text," *Bioinformatics,* vol. 34, pp. 1547-1554, 2018.

[51] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annu Rev Biomed Eng,* vol. 19, pp. 221-248, 2017.

[52] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter*, et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell,* vol. 172, pp. 1122-1131 e9, 2018.

[53] H. Teng, M. D. Cao, M. B. Hall, T. Duarte, S. Wang, and L. J. M. Coin, "Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning," *Gigascience,* vol. 7, Art. no. giy037, 2018.

[54] A. Gharehbaghi and M. Linden, "A deep machine learning method for classifying cyclic time series of biological signals using time-growing neural network," *IEEE Trans Neural Netw Learn Syst,* DOI: 10.1109/TNNLS.2017.2754294, 2017.

[55] A. Telenti, C. Lippert, P. C. Chang, and M. DePristo, "Deep learning of genomic variation and regulatory network data," *Hum Mol Genet,* vol. 27, pp. R63-R71, 2018.

[56] C. Savojardo, P. L. Martelli, P. Fariselli, and R. Casadio, "DeepSig: deep learning improves signal peptide detection in proteins," *Bioinformatics,* vol. 34, pp. 1690-1696, 2018.

[57] J. Liu, X. Wang, Y. Cheng, and L. Zhang, "Tumor gene expression data classification via sample expansion-based deep learning," *Oncotarget,* vol. 8, pp. 109646-109660, 2017.

[58] Y. Kong and T. Yu, "A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data," *Bioinformatics,* Art. no. bty429, 2018.

[59] K. Sharma, C. Rupprecht, A. Caroli, M. C. Aparicio, A. Remuzzi, M. Baust*, et al.*, "Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease," *Sci Rep,* vol. 7, Art. no. 2049, 2017.

[60] P. Jackson, N. Hardcastle, N. Dawe, T. Kron, M. S. Hofman, and R. J. Hicks, "Deep learning renal segmentation for fully automated radiation dose estimation in unsealed source therapy," *Front Oncol,* vol. 8, Art. no. 215, 2018.

[61] Y. C. Lai, C. Y. Wang, S. H. Moi, C. H. Wu, C. H. Yang, and J. B. Chen, "Factors associated with functional performance among patients on hemodialysis in Taiwan," *Blood Purif,* vol. 46, pp. 12-18, 2018.

[62] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics,* vol. 21, pp. 631-643, 2005.

[63] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, Jeffrey Dean*, et al.*, "TensorFlow: a system for large-scale machine learning," *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*, Savannah, GA, USA, 2016, pp. 265-283.

[64] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel*, et al.*, "Scikit-learn: Machine Learning in Python," *J Mach Learn Res,* vol. 12, pp. 2825-2830, 2011.

[65] V. Begin, C. Deziel, and F. Madore, "Biofeedback regulation of ultrafiltration and dialysate conductivity for the prevention of hypotension during hemodialysis," *ASAIO J,* vol. 48, pp. 312-315, 2002.

[66] W. Sulowicz and A. Radziszewski, "Pathogenesis and treatment of dialysis hypotension," *Kidney International,* vol. 70, pp. S36-S39, 2006.

[67] E. Movilli, P. Gaggia, R. Zubani, C. Camerini, V. Vizzardi, G. Parrinello*, et al.*, "Association between high ultrafiltration rates and mortality in uraemic patients on regular haemodialysis. A 5-year prospective observational multicentre study," *Nephrol Dial Transplant,* vol. 22, pp. 3547-3552, 2007.

[68] J. E. Flythe, S. E. Kimmel, and S. M. Brunelli, "Rapid fluid removal during dialysis is associated with cardiovascular morbidity and mortality," *Kidney Int,* vol. 79, pp. 250-257, 2011.

[69] K. C. W. Leung, R. R. Quinn, P. Ravani, H. Duff, and J. M. MacRae, "Randomized crossover trial of blood volume monitoring-guided ultrafiltration biofeedback to reduce intradialytic hypotensive episodes with hemodialysis," *Clin J Am Soc Nephrol,* vol. 12, pp. 1831-1840, 2017.

[70] K. H. Chu, C. P. Li, and J. Qi, "Ribosomal RNA as molecular barcodes: a simple correlation analysis without sequence alignment," *Bioinformatics,* vol. 22, pp. 1690-1701, 2006.

[71] C. Mora, D. P. Tittensor, S. Adl, A. G. Simpson, and B. Worm, "How many species are there on Earth and in the ocean?," *PLoS Biol,* vol. 9, Art. no. e1001127, 2011.

[72] P. D. Hebert, M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis, "Identification of birds through DNA barcodes," *PLoS Biol,* vol. 2, p. e312, 2004.

[73] M. Blaxter, J. Mann, T. Chapman, F. Thomas, C. Whitton, R. Floyd*, et al.*, "Defining operational taxonomic units using DNA barcode data," *Philos Trans R Soc Lond B Biol Sci,* vol. 360, pp. 1935-1943, 2005.

[74] E. Weitschek, R. Van Velzen, G. Felici, and P. Bertolazzi, "BLOG 2.0: a software system for character-based species classification with DNA Barcode sequences. What it does, how to use it," *Mol Ecol Resour,* vol. 13, pp. 1043-1046, 2013.

[75] E. Weitschek, G. Fiscon, and G. Felici, "Supervised DNA Barcodes species classification: analysis, comparisons and results," *BioData Min,* vol. 7, Art. no. 4,

2014.

[76] A. Fiannaca, M. La Rosa, R. Rizzo, and A. Urso, "A k-mer-based barcode DNA classification methodology based on spectral representation and a neural gas network," *Artif Intell Med,* vol. 64, pp. 173-184, 2015.

[77] H. Zeng and D. K. Gifford, "Predicting the impact of non-coding variants on DNA methylation," *Nucleic Acids Res,* vol. 45, Art. no. e99, 2017.

[78] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, "MEGA X: molecular evolutionary genetics analysis across computing platforms," *Mol Biol Evol,* vol. 35, pp. 1547-1549, 2018.

[79] Z. M. Fadlullah, F. X. Tang, B. M. Mao, N. Kato, O. Akashi, T. Inoue*, et al.*, "State-of-the-art deep learning: evolving machine intelligence toward tomorrow's intelligent network traffic control systems," *IEEE Commun Surv Tutor,* vol. 19, pp. 2432-2455, 2017.

[80] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: a review and new perspectives," *IEEE Trans Pattern Anal Mach Intell,* vol. 35, pp. 1798-1828, 2013.

[81] G. Aoki and Y. Sakakibara, "Convolutional neural networks for classification of alignments of non-coding RNA sequences," *Bioinformatics,* vol. 34, pp. i237-i244, 2018.

[82] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning," *Nat Biotechnol,* vol. 33, pp. 831-838, 2015.

[83] J. J. Almagro Armenteros, C. K. Sonderby, S. K. Sonderby, H. Nielsen, and O. Winther, "DeepLoc: prediction of protein subcellular localization using deep learning," *Bioinformatics,* vol. 33, pp. 3387-3395, 2017.

[84] R. Singh, J. Lanchantin, G. Robins, and Y. Qi, "DeepChrome: deep-learning for predicting gene expression from histone modifications," *Bioinformatics,* vol. 32, pp. i639-i648, 2016.

[85] Q. Liu, F. Xia, Q. Yin, and R. Jiang, "Chromatin accessibility prediction via a

hybrid deep convolutional neural network," *Bioinformatics,* vol. 34, pp. 732-738, 2018.

[86] B. Yang, F. Liu, C. Ren, Z. Ouyang, Z. Xie, X. Bo*, et al.*, "BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone," *Bioinformatics,* vol. 33, pp. 1930-1936, 2017.

[87] Z. Avsec, M. Barekatain, J. Cheng, and J. Gagneur, "Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks," *Bioinformatics,* vol. 34, pp. 1261-1269, 2018.

[88] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J Mach Learn Res,* vol. 15, pp. 1929-1958, 2014.

[89] P. D. Hebert, A. Cywinska, S. L. Ball, and J. R. deWaard, "Biological identifications through DNA barcodes," *Proc Biol Sci,* vol. 270, pp. 313-321, 2003.

[90] C. H. Yang, K. C. Wu, L. Y. Chuang, and H. W. Chang, "Decision tree algorithm-generated single-nucleotide polymorphism barcodes of rbcl genes for 38 Brassicaceae species tagging," *Evol Bioinform,* vol. 14, Art. no. 1176934318760856, 2018.

[91] C. H. Yang, K. C. Wu, H. U. Dahms, L. Y. Chuang, and H. W. Chang, "Single nucleotide polymorphism barcoding of cytochrome c oxidase I sequences for discriminating 17 species of Columbidae by decision tree algorithm," *Ecol Evol,* vol. 7, pp. 4717-4725, 2017.

Publication list

Journal paper

1. Cheng-Hong Yang, **Kuo-Chuan Wu**, Yu-Shiun Lin, Li-Yeh Chuang, and Hsueh-Wei Chang, "Protein folding prediction in the HP model using ions motion optimization with a greedy algorithm", *BioData Mining,* 2018. (accepted at 2018.07.22) **[SCI]**

2. Cheng-Hong Yang, **Kuo-Chuan Wu**, Li-Yeh Chuang, and Hsueh-Wei Chang, "Decision Tree Algorithm–Generated Single-Nucleotide Polymorphism Barcodes of *rbcL* Genes for 38 Brassicaceae Species Tagging," *Evolutionary Bioinformatics*, vol. 14, Art. no. 1176934318760856, 2018. **[SCI]**

3. Cheng-Hong Yang, **Kuo-Chuan Wu**, Hans-Uwe Dahms, Li-Yeh Chuang, and Hsueh-Wei Chang, "Single nucleotide polymorphism barcoding of cytochrome c oxidase I sequences for discriminating 17 species of Columbidae by decision tree algorithm," *Ecology and Evolution,* vol. 7, pp. 4717-4725. **[SCI]**

4. Cheng-Hong Yang, Yu-Shiun Lin, Sin-Hua Moi, **Kuo-Chuan Wu**, Li-Yeh Chuang, and Hsueh-Wei Chang, "Hybrid high exploration particle swarm optimization algorithm improves the prediction of the 2-dimensional hydrophobic-polar model for protein folding," *Current Bioinformatics,* vol. 12, pp. 1-11, 2017. **[SCI]**

5. Cheng-Hong Yang, **Kuo-Chuan Wu**, and Li-Yeh Chuang, "Breast Cancer Risk Prediction Using Ions Motion Optimization Algorithm," *Journal of Life Sciences and Technologie,* vol. 4, pp. 49-55, 2016.

6. Li-Yeh Chuang, Cheng-San Yang, **Kuo-Chuan Wu**, Hsueh-Wei Chang, and Cheng-Hong Yang, "Hybrid Taguchi and Binary Particle Swarm Optimization Method for Tumor Classification," *International Journal of Cancer Research and Prevention,* vol. 5, pp. 133-151, 2012.

7. Li-Yeh Chuang, Cheng-SanYang, **Kuo-Chuan Wu**, and Cheng-HongYang, "Gene selection and classification using Taguchi chaotic binary particle swarm optimization," *Expert Systems with Applications,* vol. 38, pp. 13367-13377, 2011. **[SCI]**

8. Li-Yeh Chuang, Cheng-Huei Yang, **Kuo-Chuan Wu**, and Cheng-Hong Yang, "A hybrid feature selection method for DNA microarray data," *Computers in Biology and Medicine,* vol. 41, pp. 228-237, 2011. **[SCI]**

Conference paper

1. Cheng-Hong Yang, **Kuo-Chuan Wu**, and Li-Yeh Chuang, "Breast Cancer Risk Prediction using Ions Motion Optimization Algorithm," in *5th International Conference on Bioinformatics and Biomedical Science*, Bali, Indonesia, 2016, pp. 52-58.

2. Cheng-Hong Yang, **Kuo-Chuan Wu**, and Li-Yeh Chuang, "Ions Motion Optimization Algorithm for SNP Epistasis Detection in Oral Cancer Risk Prediction," in *The 6th International Conference on Engineering and Applied Sciences (ICEAS 2016)*, Hong Kong, 2016, pp. 16-24.

3. Cheng-Hong Yang, Li-Yeh Chuang, Hsueh-Wei Chang, and **Kuo-Chuan Wu**, "Support vector machine-based prediction for oral cancer using four SNPs in DNA repair genes," in *International MultiConference of Engineers and Computer Scientists*, Hong Kong, 2011, pp. 426-429.

4. 楊正宏、**吳國銓**、莊麗月、張學偉，"以決策理論為基礎用於 DNACOI 物種編碼"，*第十七屆離島資訊技術與應用研討會*，澎湖，2018 年 5 月，293-299。