Accident Analysis and Prevention 153 (2021) 106034

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Single-vehicle crash severity outcome prediction and determinant extraction using tree-based and other non-parametric models

Xintong Yan, Jie He *, Changjian Zhang, Ziyang Liu, Boshuai Qiao, Hao Zhang

School of Transportation, Southeast University, 2 Si pai lou, Nanjing, 210096, PR China

ARTICLE INFO

Handled by E. Dahlen

Keywords:
Single-vehicle accident analysis
Accident severity
Severity prediction
Non-parametric model
Tree-based models

ABSTRACT

Single-vehicle crashes are more fatality-concentrated and have posed increasing challenges in traffic safety, which is of great research necessity. Tremendous previous studies have conducted relevant analysis with econometric modeling approaches, whereas the ability of non-parametric methods to predict crash severity is still smattering of knowledge. Consequently, the main objective of this paper is to conduct single-vehicle crash severity prediction with different tree-based and non-parameter models. An alternate aim is to identify the intrinsic mechanism of how contributing factors determine single-vehicle crash severity. By virtue of Grid-Search method, this paper conducted fine-tuning of different models to obtain the best performances based on five crash severity sub-datasets. For model evaluation, the accuracy indicators were calculated in training, validation and test sets, respectively. Besides, feature importance extraction was undertaken based on the results of model comparison. The finding indicated that these models didn't exhibit a huge performance difference for crash severity prediction in the same severity level; however, the performances of the models did vary among different datasets, with an average training accuracy of 99.27 %, 96.4 %, 86.98 %, 86.84 %, 71.76 % in fatal injury, severe injury, visible injury, complaint of pain, PDO crash datasets, respectively. Additionally, it was found that in each severity dataset, the indicator *urban freeways* is a determinant factor that leads to the occurrence of crashes while *rural freeways* is more related to more severe crashes (i.e., fatal and severe crashes). This paper can provide valuable information for model selection and tuning in accident severity prediction. Future research could consider the influences that temporal instability of contributing features has on the model performances.

1. Introduction

Single-vehicle crashes are frequently associated with a disproportionate number of serious and fatal crashes, which pose increasing challenges in traffic safety (Wu et al., 2016; Yu et al., 2019). As reported, in the United States, single-vehicle accounted for only about 16 % of all traffic accidents in 2018, while leading to 36.9 % of motor vehicle crash fatalities (National Highway Traffic Safety Administration, 2019), indicating that single-vehicle crashes are more fatality-concentrated than multi-vehicle crashes (Yu et al., 2019) and there is a great necessity to develop more powerful and efficient instruments for single-vehicle accident analysis, prediction and prevention.

Given the difference of internal mechanism between single-vehicle and multi-vehicle crashes (Martensen and Dupont, 2013; Wu et al., 2013; Wu et al., 2016; Yu et al., 2019), substantial researches have been conducted in an effort to separately investigate significant causal factors and their impacts on injury severities, among which the unique characteristics and attributes associated with single-vehicle crashes have been explored extensively (e.g., Behnood and Mannering, 2017, 2015; Chen et al., 2016; Feng et al., 2016; Gong and Fan, 2017; Li et al., 2018; Wu et al., 2013).

Regarding single-vehicle crash severity analyses, econometric modeling approaches such as mixed logit, nested logit, latent class were frequently utilized for better understanding of unobserved heterogeneity among crash factor variables, for instance, Li et al. (2019b) and Yu et al. (2019) both developed latent class mixed logit models to investigate highway single-vehicle crashes and the effects of significant contributing factors to driver injury severity, consequently demonstrating that some fixed-effect variables such as wet, overturning, snow had significant impacts on different levels of injury severity outcomes. Wu et al. (2016) utilized nested logit and mixed logit models to account for the correlation between severity categories, which revealed similar

* Corresponding author.
E-mail addresses: 230198699@seu.edu.cn (X. Yan), hejie@seu.edu.cn (J. He), 230189680@seu.edu.cn (C. Zhang), lzyseu6257@outlook.com (Z. Liu), 230199154@seu.edu.cn (B. Qiao), andyhao@seu.edu.cn (H. Zhang).

https://doi.org/10.1016/j.aap.2021.106034
Received 22 November 2020; Received in revised form 7 January 2021; Accepted 7 February 2021
Available online 26 February 2021

Table 1
Descriptive statistics.

	SEV0		SEV1		SEV2		SEV3		SEV4		total
	frequency	%	frequency	%	frequency	%	frequency	%	frequency	%	
Time Characteristics											
Day											
Daytime(10:00−16:59)	6148	63.59	116	1.20	307	3.18	1457	15.07	1640	16.96	9668
Night(20:00−next 6:59)	10,259	64.60	189	1.19	667	4.20	2569	16.18	2197	13.83	15,881
Peak Hours(7:00−10:00;17:00−20:00)	4995	64.22	83	1.07	246	3.16	1206	15.51	1248	16.05	7778
Week											
Monday	3566	61.36	90	1.55	248	4.27	1057	18.19	851	14.64	5812
Tuesday	3025	65.53	42	0.91	133	2.88	689	14.93	727	15.75	4616
Wednesday	2669	66.21	37	0.92	141	3.50	580	14.39	604	14.98	4031
Thursday	2691	65.52	45	1.10	137	3.34	565	13.76	669	16.29	4107
Friday	2933	65.22	37	0.82	121	2.69	669	14.88	737	16.39	4497
Saturday	3320	65.32	62	1.22	186	3.66	766	15.07	749	14.74	5083
Sunday	3198	61.73	75	1.45	254	4.90	906	17.49	748	14.44	5181
Environmental Characteristics											
Weather											
Clear	15,263	63.97	304	1.27	939	3.94	3965	16.62	3389	14.20	23,860
Bad(Cloudy/Raining/Snowing/Wind)	6139	64.85	84	0.89	281	2.97	1267	13.38	1696	17.91	9467
Road Surface											
Dry	16,801	63.89	336	1.28	1035	3.94	4380	16.66	3746	14.24	26,298
Wet	4165	65.24	48	0.75	162	2.54	769	12.05	1240	19.42	6384
Snowy/Icy/ Slippery/Muddy	436	67.60	4	0.62	23	3.57	83	12.87	99	15.35	645
Light Condition											
Daylight	10,298	63.55	184	1.14	491	3.03	2507	15.47	2725	16.82	16,205
Dusk-Dawn	855	63.66	12	0.89	52	3.87	199	14.82	225	16.75	1343
Dark-Street Lights	4254	64.94	65	0.99	250	3.82	1109	16.93	873	13.33	6551
Dark- No Street Lights	5995	64.97	127	1.38	427	4.63	1417	15.36	1262	13.68	9228
Roadway Characteristics											
Roadway Classification											
Urban Freeways	12,791	65.54	165	0.85	601	3.08	2890	14.81	3070	15.73	19,517
Urban Two Lane Roads	566	63.67	9	1.01	36	4.05	141	15.86	137	15.41	889
Urban Multilane Non-freeways	913	65.35	16	1.15	51	3.65	209	14.96	208	14.89	1397
Rural Freeways	2511	63.03	72	1.81	178	4.47	609	15.29	614	15.41	3984
Rural Two Lane Roads	3131	59.71	89	1.70	274	5.23	1013	19.32	737	14.05	5244
Rural Multilane Non-freeways	1213	65.71	35	1.90	60	3.25	292	15.82	246	13.33	1846
Other	277	61.56	2	0.44	20	4.44	78	17.33	73	16.22	450
Roadway Condition											
Holes, Deep Ruts	66	80.49	1	1.22	1	1.22	5	6.10	9	10.98	82
Loose Material on Road	146	83.43	1	0.57	3	1.71	14	8.00	11	6.29	175
Obstruction on Roadway	431	86.55	1	0.20	4	0.80	28	5.62	34	6.83	498
Construction-Repair Zone	418	72.44	7	1.21	17	2.95	71	12.31	64	11.09	577
Reduced Road Width	16	64.00	0	0.00	0	0.00	3	12.00	6	24.00	25
Flooded	63	57.80	0	0.00	1	0.92	12	11.01	33	30.28	109
G-Other	73	63.58	0	1.19	2	3.75	12	16.02	21	15.45	108
H-No Unusual Conditions	20,189	63.58	378	1.19	1192	3.75	5087	16.02	4907	15.45	31,753
Accident Characteristics											
Accident type											
Head-on	169	0.79	2	0.52	11	0.90	39	0.75	27	0.53	248
Sideswipe	232	1.08	0	0.00	6	0.49	44	0.84	58	1.14	340
Rear End	27	0.13	0	0.00	1	0.08	5	0.10	7	0.14	40
Broadside	20	0.09	1	0.26	1	0.08	3	0.06	6	0.12	31
Hit Object	17,995	84.08	293	75.52	891	73.03	3871	73.99	4103	80.69	27,153
Overturned	1745	8.15	85	21.91	294	24.10	1199	22.92	787	15.48	4110
Auto-Pedestrian	1	0.00	0	0.00	0	0.00	0	0.00	3	0.06	4
Other	1213	5.67	7	1.80	16	1.31	71	1.36	94	1.85	1401
Collision Location											
Beyond Median/Barrier	83	64.84	3	2.34	7	5.47	18	14.06	17	13.28	128
Beyond Shoulder - Driver's Left	6815	58.40	108	0.93	476	4.08	2064	17.69	2206	18.90	11,669
Left Shoulder Area	31	50.00	5	8.06	1	1.61	16	25.81	9	14.52	62
Left Lane	1237	85.08	6	0.41	22	1.51	107	7.36	82	5.64	1454
Interior Lanes	1542	89.60	3	0.17	13	0.76	90	5.23	73	4.24	1721
Right Lane	2137	83.90	7	0.27	33	1.30	192	7.54	178	6.99	2547
Right Shoulder Area	150	59.06	4	1.57	9	3.54	44	17.32	47	18.50	254
Beyond Shoulder - Driver's Right	8616	59.89	246	1.71	620	4.31	2547	17.70	2357	16.38	14,386
Gore Area	317	68.17	2	0.43	17	3.66	72	15.48	57	12.26	465
Other	474	73.95	4	0.62	22	3.43	82	12.79	59	9.20	641
Vehicle Characteristics											
Vehicle Type											
Truck	1281	76.94	20	1.20	41	2.46	185	11.11	138	8.29	1665
Passenger Car	17,180	63.60	300	1.11	962	3.56	4249	15.73	4321	16.00	27,012
Pickup	2734	62.36	68	1.55	207	4.72	777	17.72	598	13.64	4384
Bus	30	57.69	0	0.00	3	5.77	8	15.38	11	21.15	52
Movement Preceding Accident											
Slowing/Stopping	197	75.48	1	0.38	6	2.30	28	10.73	29	11.11	261

(*continued on next page*)

Table 1 (*continued*)

	SEV0		SEV1		SEV2		SEV3		SEV4		total
	frequency	%	frequency	%	frequency	%	frequency	%	frequency	%	
Proceeding Straight	8483	76.12	45	0.40	209	1.88	1041	9.34	1366	12.26	11,144
Ran Off Road	3311	55.97	138	2.33	348	5.88	1171	19.79	948	16.02	5916
Backing	95	95.00	0	0.00	0	0.00	3	3.00	2	2.00	100
Passing Other Vehicle	26	40.00	3	4.62	7	10.77	18	27.69	11	16.92	65
Changing Lanes	286	64.13	4	0.90	14	3.14	64	14.35	78	17.49	446
Turning	2244	60.42	43	1.16	151	4.07	712	19.17	564	15.19	3714
Crossed Into Opposing Lane	25	46.30	5	9.26	3	5.56	16	29.63	5	9.26	54
Merging	43	82.69	0	0.00	0	0.00	3	5.77	6	11.54	52
Other	6692	57.81	149	1.29	482	4.16	2176	18.80	2076	17.94	11,575
Total	21,402		388		1220		5232		5085		33,327

results in terms of identifying contributing factors for driver injury severities. To test temporal stability of factors affecting driver-injury severities in single-vehicle crashes, Behnood and Mannering (2015) conducted annual mixed logit models with a nine-year crash dataset in terms of multiple variables affecting injury severities including driver-contributing factors, location and time of day, environmental conditions, etc.

In addition, ordered logit and probit (e.g., Fountas et al., 2020; Naik et al., 2016; Osman et al., 2018) were also prevailing in single-vehicle crashes analyses, e.g., Fountas et al. (2020) employed a zero-inflated hierarchical ordered probit approach with correlated disturbances to account for the joint effect of weather and lighting conditions on injury severities of single-vehicle accidents. Likewise, to understand weather impacts on single-vehicle truck crash injury severity, random parameters ordered logit was estimated by Naik et al. (2016), with the result showing that integrating comprehensive weather data with crash data provided useful insights into factors associated with single-vehicle truck crash injury severity.

These previous statistical methods have provided comprehensive insights into contributing factors of crash severities in single-vehicle crashes. Besides, owing to being established on the basis of rigorous mathematical hypothesis and functional formations, these statistical methods have the capability to explain the mathematical relation between crash severities and contributing factors (Tang et al., 2020). However, due to the limitations that the assumptions for data distribution and the pre-defined underlying relationships have to be satisfied (Xing et al., 2020; Xu et al., 2013), traditional statistical framework tends to perform less satisfactorily in target prediction than its counterpart machine learning (ML) method, which can serve as another robust and powerful tool for accident analysis, especially in the field of accident prediction.

Past work has obtained marvelous achievements in accident analysis with ML approaches, of which Random Forests (RF), K-Nearest Neighbor (KNN), Decision Trees (DT), Support Vector Machines (SVM), Multilayer Perception (MLP), Gradient Boosting Decision Tree (GBDT) are extensively employed (e.g., Alkheder et al., 2017; Ding et al., 2017; Li et al., 2020; Shi and Abdel-aty, 2015; Tang et al., 2019; Yu et al., 2016). For instance, Li et al. (2020) utilized a RF classifier to analyze the correlation between the dependent and the independent variables. Xing et al. (2020) compared five typical ML models including KNN, ANN, SVM, DT, and RF to examine the relationship between influencing factors and vehicle collision risk. By conducting comparisons between four statistical and ML methods including Multinomial Logit (MNL), KNN, SVM and RF, Iranitalab and Khattak (2017) found that KNN had the best prediction performance in overall and in more severe crashes.

Despite abundant studies relating to accident analysis with ML approaches, in crash severity prediction domain, especially for single-vehicle crash severity, the relevant application and practice is marginally monotonous, dominated by tree-based models (e.g., Abellán et al., 2013; Chang and Chien, 2013; Das et al., 2009; Kashani et al., 2014). Other outstanding classifiers, such as Quadratic Discriminant Analysis (QDA), MLP, KNN have not been adequately applied in crash severity prediction. Additionally, due to lacking comparisons of performances of the aforementioned non-parametric models in crash severity prediction, it is still implicit in their abilities to implement accurate prediction. Likewise, sample distribution in different levels of crash severity can be imbalanced, indicating that even with the same ML method, the prediction accuracy in different crash severities can be of huge difference, which was not discussed in previous studies.

Thus, to address these research gaps and facilitate the construction of a more explicit single-vehicle crash severity analysis framework with the current advanced ML models, the objectives of this paper are as follows: (1) to conduct crash severity prediction using different tree-based and non-parametric models; (2) to test the hypothesis that the generalization ability may vary in different models of predicting different levels of crash severity due to the unbalanced sample distribution (3) and to further extract critical features that have significant impacts on determining different single-vehicle severities with the best models on the basis of model performance evaluation in the first phase.

2. Data preprocessing and description

2.1. Data preprocessing

Data records were obtained for all single-vehicle crashes occurring in California from January 1, 2017 until December 31, 2017, which were provided by Highway Safety Information System (HSIS), managed by the University of North Carolina Highway Safety Research Center under a contract with the Federal Highway Administration (FHWA), 2020. Crash information used in this articile was recorded in two subfiles as follows:

(1) Accident Sub-file: contains more than 40 basic variables describing the overall crash (i.e., time and location, weather, lighting, collision severity, accident type, etc.).
(2) Vehicle Sub-file: contains more than 30 specific variables related to contributing factors, object struck, and movements of each vehicle involved in the collision

To compile the information obtained by the separate files into one dataset, the unique case number labeling each crash and its involved vehicles was used as a linkage for each crash (from the accident files) and each vehicle involved in that collision (from the vehicle files). To ensure the validation and reliability of dataset, there is a great necessity to conduct feature engineering. Thus, the joint crash dataset was preprocessed to remove some features and records with missing information. Through the feature selection and data cleaning, a total of 3937 crash samples were removed with 33,327 being retained. Table 1 summarizes all the variables' descriptive statistics. Additionally, to transform the original dataset into the appropriate form that can be understood by ML language, one-hot coding (OHC) method (Arnold and Morgan, 2019; Taleshmekaeil et al., 2012) was utilized to convert the original nominal variables to vectorized metrics consisting of dichotomous elements. To be more precisely, take the light condition related

variables as the example, *Daylight, Dusk-Dawn, Dark-Street Lights,* and *Dark-No Street Lights* can be denoted as [1, 0, 0, 0], [0, 1, 0, 0], [0, 0, 1, 0] and [0, 0, 0, 1], respectively. Besides, given the rare event feature of injury crashes, Synthetic Minority Over-sampling Technique (SMOTE) was adopted to deal with the imbalanced leaning issue. SMOTE is a typical re-sampling technique proposed by Chawla et al., (2002) that has been considered to be one of the most powerful re-sampling algorithms with the basic idea to produce synthetic instances from the minor class.

2.2. Data description

As presented in Table 1, the crash severity can be categorized as five scales: 0- property damage only (PDO); 1-fatal injury; 2-severe injury; 3-visible injury; 4-complaint of pain. Of the 33,327 crashes, 64.22 % resulted in PDO, 1.16 % resulted in fatal injury, 3.66 % involved severe injury, 15.70 % involved visible injury and 15.26 % resulted in complaint of pain.

Regarding time characteristics of the crashes, nominal features were divided from two dimensions: day and week, and it was found that in the Day dimension, night tended to witness more accidents, whereas in the Week dimension, accidents were less likely to occur in the midweek. For environmental characteristics, it could not be obviously observed that adverse weather or road surface condition will increase the frequency and severity of accidents as believed in some previous studies (e.g., Jung et al., 2014; Kim et al., 2013). One possible explanation is that driver might be more cautious when driving on worse driving environments (Haque et al., 2012; Zhu and Srinivasan, 2011). Compared to dark conditions with street lights, those without street lights were more likely to trigger fatal crashes. With regard to roadway characteristics, rural two lane roads were found to have a higher risk of more crashes relative to the other rural roadways. In terms of accident characteristics, collisions beyond shoulders seemed to be more frequent, with the total amount reaching 15,431, accounting for nearly half of the total crashes. As for the vehicle characteristics, passenger cars brought out most of the crashes (81 %), with the total amount (17,180) being about 5 times larger than the pickup crashes (2734), the second leading crashes classified by vehicle types.

3. Method

3.1. Modeling approach

3.1.1. Tree-based models

3.1.1.1. Decision trees (DT). DT model is a non-parametric supervised learning method used for classification and regression, with the goal to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. DT model is characterized with desirable interpretability and visualization, with low requirements for data preparation, which overwhelms other ML methods and has been extensively used in accident analyses (e.g., Abellán et al., 2013; Alkheder et al., 2020; Figueira et al., 2017). Commonly used DT algorithms include ID3, C4.5, C5.0 and CART (Classification and Regression Trees).

In this study, CART (Breiman et al., 1984) is used to construct crash severity classification trees using the crash-related features and thresholds that yield the largest information gain at each node. Larger information gain represents higher impurity of the node, indicating a better splitting effect. Commonly, impurity can be measured by Gini index, which can be defined as:

$$\mathrm{Gini}(\mathrm{X}_m) = \sum_k p_{mk}(1 - p_{mk}) \tag{1}$$

where X_m is the training data at node m, p_{mk} is the possibility that class k occurs at node m.

3.1.1.2. Random forests (RF). RF (Breiman, 2001) is a typical assembling ML methods derived from tree-based models, where each tree in the ensemble is built from a sample drawn with replacement from the training set. RF can run efficiently on large databases and handle thousands of input variables without variable deletions. Besides, it is unexcelled in accuracy among current algorithms. The essence of RF is to construct a robust vote classifier by combining several base estimators to maximize their strengths on the basis of "Bagging" (Bootstrap Aggregating) theorem. The effect of RF classification is related to two major factors: correlation of any two trees and classification ability of each tree in the forest. Generally, to obtain better model performance, the only hyper-parameter needed to be tuned in RF is the number of clusters, which is also conducted in this paper.

3.1.1.3. Adaptive boosting (AdaBoost). Boosting is a prevailing algorithm in machine learning in recent years, among which AdaBoost (Freund and Schapire, 1995) can be easily applied to practical problems owing to its high efficiency, accuracy and generalization. The core principle of AdaBoost is to fit a sequence of weak learners on repeatedly modified versions of the samples. The predictions from all of them are then combined through a weighted majority vote to produce the final prediction. Different from other boosting classifiers, AdaBoost doesn't depend on any prior information related to the voter. Instead, it is adjusting for weights of the voter adaptively, which can be determined as (Hongpu and Beiji, 2020):

$$\mathrm{H}(\mathrm{x}) = \begin{cases} 1, \sum_{t=1}^{k} \alpha_t h_t(x) > \theta \\ -1, otherwise \end{cases} \tag{2}$$

where, x represents the sample, $\mathrm{H}(\mathrm{x})$ is the voter, α_t is the weight of the base estimator $h_t(x)$, k is the number of the estimators, θ is the threshold determined by $\sum_{t=1}^{k} \alpha_t$.

3.1.1.4. Gradient boosting decision tree (GBDT). GBDT is also a boosting algorithm to connect multiple base estimators. It is different from AdaBoost since it is based on the Gradient Descent theory to optimize arbitrary differentiable loss functions, i.e., the base estimators were trained in the negative gradient direction of the loss function and every time a new base estimator is added, its target is to minimize the loss function. For GBDT model, the major adjustable hyper-parameters are the number of estimators and learning rate. Normally, the decrese of learning rate can ameliorate the generation ability of the model, indicating that the model is of high repurposability and consequently obtain better performance in untrained test dataset, which is quite important in ML framework (Chollet, 2018).

3.1.1.5. Extreme gradient boosting (XGBoost). XGBoost (Chen and Guestrin, 2016) is an improved model based on GBDT via integrating block and regularization idea to enhance the accuracy and speed of the model, as well as to prevent overfitting (i.e., the trained model performs well in training set while not in the test set). Compared to GBDT, XGBoost has been reported to have the following merits:(1) XGBoost integrates the complexity of the tree model to the regularization term to avoid over fitting, and thus its generalization ability outperforms that of GBDT. (2) The loss function of XGBoost is expanded by Taylor expansion with the first and the second derivative to speed up the optimization. (3) GBDT only supports CART as the base classifier, while XGBoost can also utilize linear classifiers through L1 and L2 regularization. Note that from a theoretical perspective, XGBoost might of be a slightly overwhelming advantage compared to GBDT, while in practice, GBDT might be more available considering its lower requirement for the hardware equipment.

To be precisely, the core idea of XGBoost is to add a penalty term to minimize the following objective:

$$\mathscr{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) \tag{3}$$

where x_i is the i-th sample, y_i is the label of the i-th sample, $\hat{y}_i^{(t-1)}$ is the prediction of i-th sample at the (t-1)-th iteration, f_t is the penalty term.

3.1.2. Non-tree-based models

3.1.2.1. Quadratic discriminant analysis (QDA). QDA is a classic classifier with closed-form solutions that can be easily computed, which has proven to work well in practice (e.g., Mahmodi et al., 2019; Menaka et al., 2014). In QDA model, it is assumed that measurements from each class are normally distributed (Muhammad et al., 2014), with the aim to project the input data to a linear subspace consisting of the directions which maximize inter-class separation while minimizing the intra-class variance.

Derived from probabilistic models which model the class conditional distribution of the data $P(X|y=k)$ for each class k, QDA can achieve label prediction by using Bayes' rule and multivariate Gaussian distribution. For each training sample $x \in R^d$, the posterior probability is represented as:

$$P(y=kx) = \frac{\frac{1}{2\pi^{\frac{d}{2}}|\sum_k|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-u_k)^t \sum_k^{-1}(x-u_k)\right)\pi_k}{\sum_l \frac{1}{2\pi^{\frac{d}{2}}|\sum_{k'}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-u_l)^t \sum_l^{-1}(x-u_l)\right)\pi_l} \tag{4}$$

where class k is chosen to maximize the posterior probability, l is the number of classes, d is the number of features, π is the priori probability that k is divided into different classes, u_k and $\sum_k$ are the mean vector and covariance of Gaussian curves belonging to different classes, respectively; u_l and $\sum_l$ are the mean vector and covariance of l classes.

3.1.2.2. Support vector machines (SVM). SVM is a supervised learning model used for classification, regression and outliers detection, with the essence to find the best separation of a hyperplane having the largest distance to the closest data point of any class (Xing et al., 2020). SVM has a good ability in dealing with high-dimensional data, even in cases where number of dimensions is greater than the number of samples. Besides, SVM is of versatility owing to its multiple choices for the specification of kernel functions. In this study, Support Vector Classifier (SVC) was mainly used for crash severity prediction, which could be described as an optimization problem with the aim to solve:

$$\xi_i \geq 0, i = 1, \ldots, n \tag{5}$$

where, x_i is the sample features, y_i represents the sample label, C is the penalty term for the sample being misclassified or within the margin boundary, ξ_i is the distance that samples are from their correct margin boundary, w and b are the training target for obtaining the best classification performance.

3.1.2.3. K-Nearest neighbors (KNN). KNN is a type of instance-based learning based on the nearest neighbors of each query point assigned the data class which has the most representatives within the nearest neighbors of the point. Normally, the distances between two points can be determined by Euclidean distance (Iranitalab and Khattak, 2017):

$$d(x_i, x_j) = \left(\sum_{k=1}^{n}(x_{ik} - x_{jk})^2\right)^{\frac{1}{2}} \tag{6}$$

where x_{ik} and x_{jk} denote the values of the k variable for observations i and j, respectively.

3.1.2.4. Bernoulli Naïve Bayes (NB). The basic idea of NB method is to apply Bayes' theorem with the "naïve" assumption of conditional independence between every pair of features given the value of the class variable (Zhang and Su, 2008), which has been worked quite well in many real-world situations. Compared to other sophisticated methods, NB classifier is time-saving since it can be extremely fast calculated. In this study, since the dataset has been divided into several sub-datasets to regard the crash severity prediction task as a dichotomous problem, Bernoulli NB is used, which is defined as:

$$P(x_i y) = P(iy)x_i + (1 - P(i_{/y}))(1 - x_i) \tag{7}$$

where $P(x_i y)$ represents the propobility that sample i belong to class y, $P(i\backslash y)$ represents the frequency that sample i occur in class y.

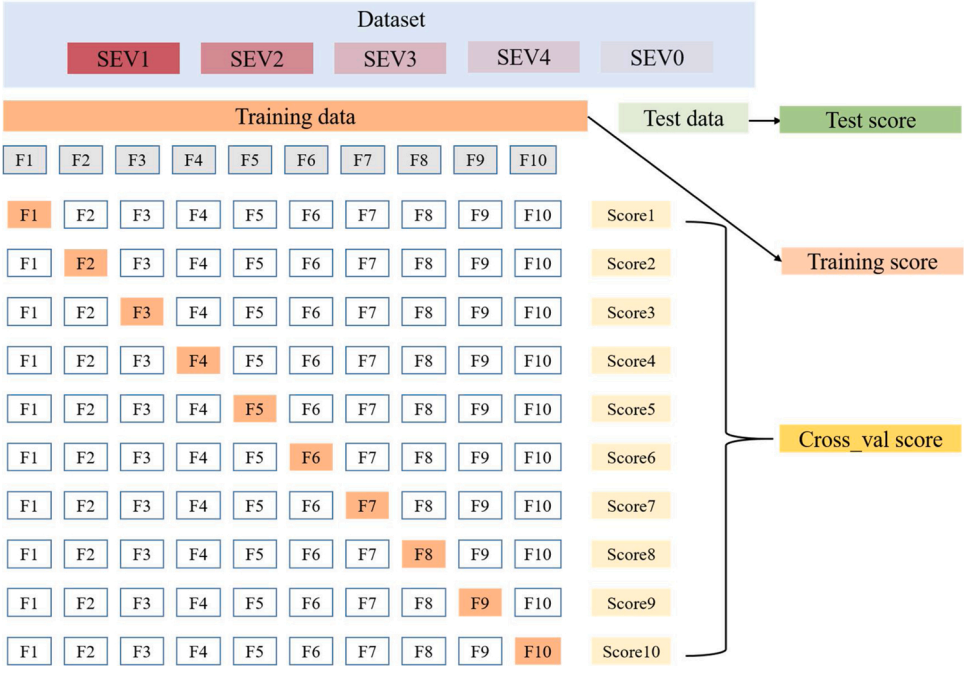


Fig. 1. Crash severity analysis dataset spilt.

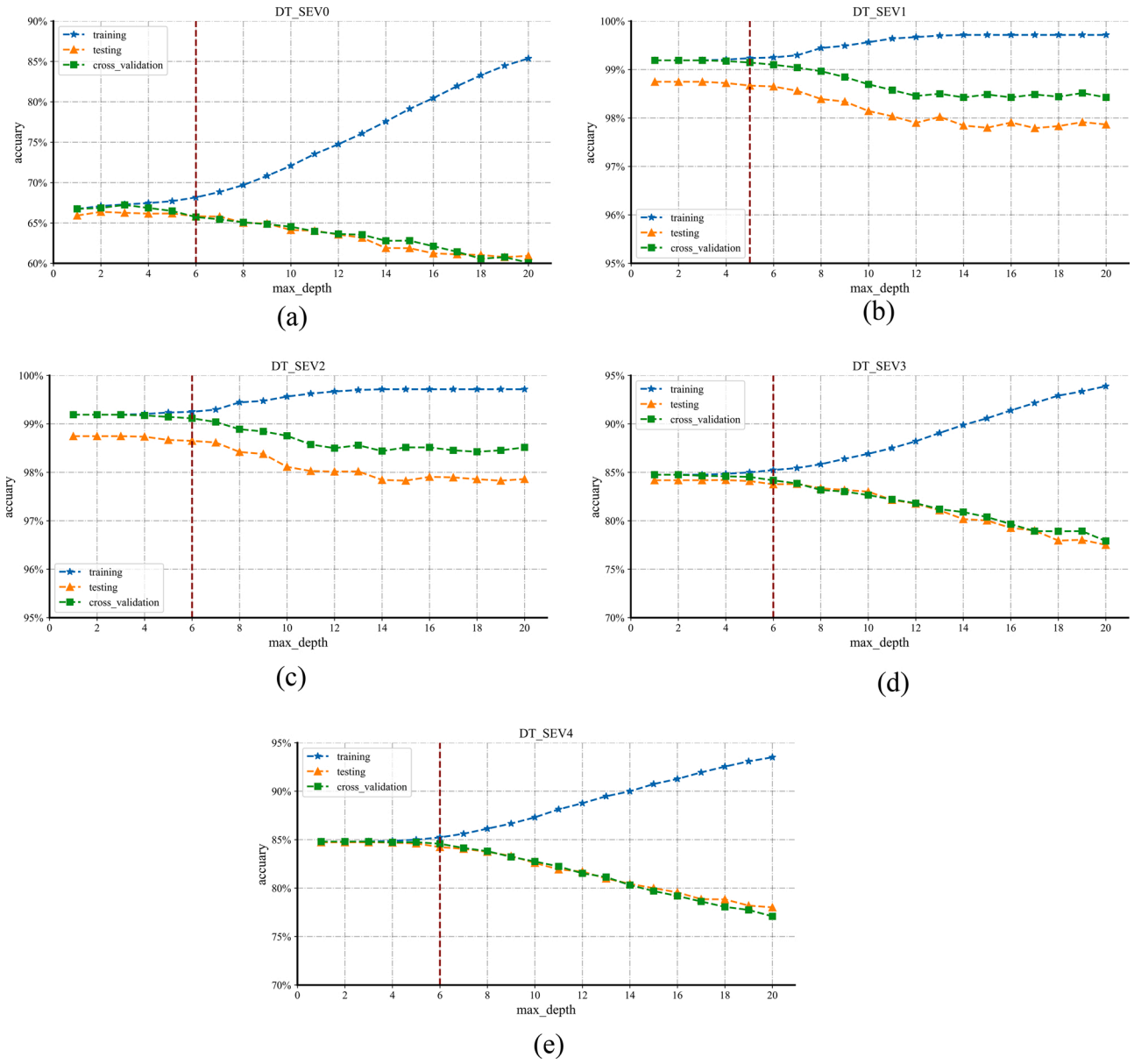


Fig. 2. Hyperparamter tuning in DT:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

3.1.2.5. Multi-layer perception (MLP). MLP, also known as Artificial Neural Network (ANN), is capable of dealing with non-linear models, and has been developed with multiple derivative versions including Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Generative Adversarial Network (GAN), etc. MLP processes non-linear problems by means of passing the input features and associated weight matrixes of different neurons belonging to the hidden layers to output the outcomes. Based on the highly interconnected neural network framework, MLP has the ability to model complex relationships between inputs and outputs (Yu et al., 2016). The output in an MLP is defined as:

$$y = f(w^T x_i + b) \tag{8}$$

where x_i is the input vector, y is the output variable; w^T represents the weight vector, b denotes the error term, and i is the number of input features.

Generally, activation function should be determined for the output layer to transform the received values from the last hidden layer into appropriate output values in a MLP model. Commonly used activation function include Softmax, Sigmoid, Relu, Tanh, etc, among which Sigmoid function tends to perform well in the dichotomous problems. This paper also utilized Sigmoid function for crash severity determination, which is defined as:

$$y(z) = \frac{1}{1 + e^{-Z}} \tag{9}$$

where y is the output value, Z is the input value from the last hidden layer.

3.2. Dataset spilt and model performance evaluation

In this study, the crash severity was considered as training label, and according to different levels of crash severity, five dichotomous labels were further determined as:

$$y_i = \begin{cases} 1, & \text{if the sample belongs to i} \\ 0, & \text{otherwise} \end{cases}, \quad i = SEV0, SEV1, SEV2, SEV3, SEV4 \tag{10}$$

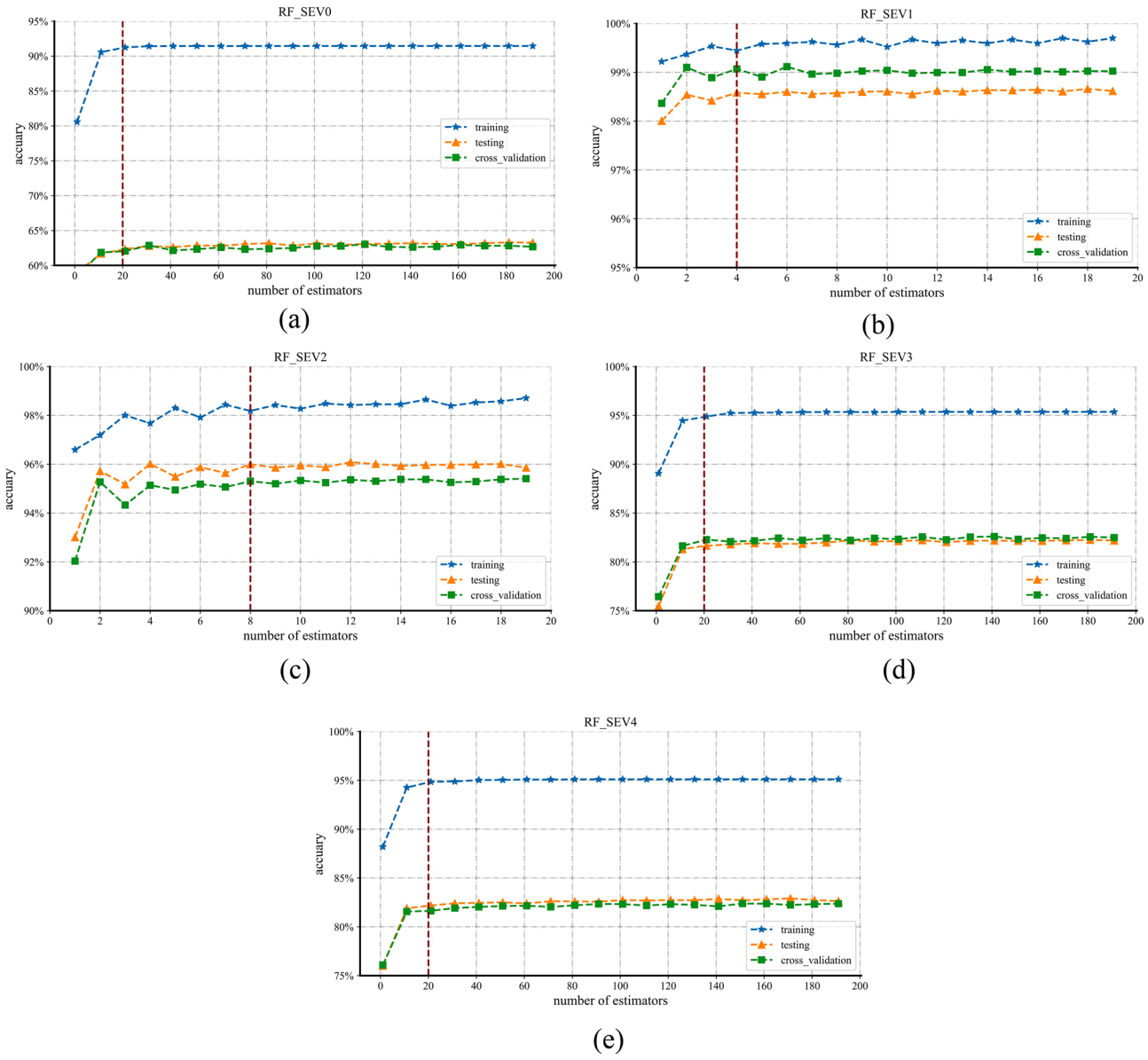


Fig. 3. Hyperparamter tuning in RF:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

By means of Eq. (10), five separate datasets were established. Note that each dataset has 71 features in the horizontal dimension and 33,327 samples in the vertical dimension.

To test the generalized prediction ability of the different models, each dataset was randomly spilled into a training set and test set with a rate of 8:2. Besides, a 10-fold cross-validation method (Xing et al., 2020) was used for further dividing the training set into a training set and validation set to check if the trained model is overfitting.

As Fig.1 presents, the performances of ten different ML models on five sub-datasets were evaluated based on three scores: training score, cross-validation score and test score. Note that although there are multiple metrics for classifier evaluation such as ROC, AUC, recall rate, precision rate, in this research, prediction accuracy is used as the evaluating metric owning to its representativeness and university in unbalanced classification problems, and the three scores were mainly represented by the accuracy, which can be calculated as:

$$\mathrm{Acc} = \frac{TP + TN}{P + N} \tag{11}$$

where $TP + TN$ is the samples that are correcly classified, $P + N$ represents all samples.

3.3. Hyper-parameter tuning and feature importance extraction

To improve the performances of the models and narrow the error gap between avoidable human-level error and Bayes' error (i.e., unavoidable machine-level bias), there is a great necessity to conduct fine-tuning engineering, i.e., to constantly adjusting the hyper-parameters of the models for maximizing the reachable performances of the model. And in this study, this step is mainly achieved by means of Grid-Search method to find the best hyperparameters of these ten models.

Additionally, to explore the intrinsic relationship between crash severity and the contributing factors, feature importance extraction is conducted on the basis of obtaining the best model in each severity dataset with the best parameters.

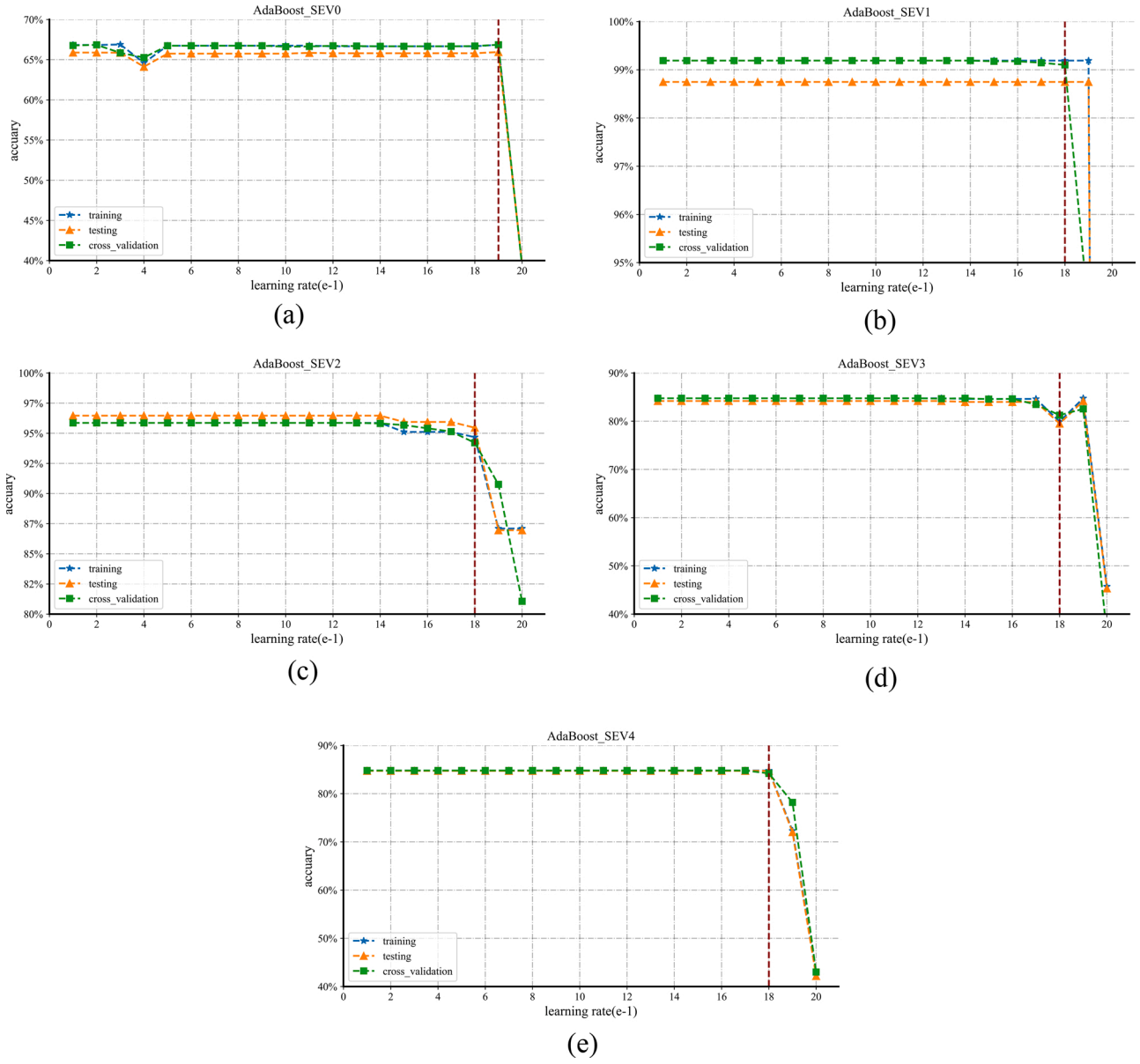


Fig. 4. Hyperparamter tuning in AdaBoost:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

4. Results and discussion

4.1. Results of hyperparamter tuning

In this research, ten ML models were applied to predict single-vehicle crash severity based on five sub-datasets divided into training sets, cross-validation sets, test sets. To narrow the gap between the avoidable human-level error and inevitable Bayes' error, Grid-Search method was utilized to tune the hyper-parameters, including regularization parameter in QDA, gamma in SVM, number of neighbors in KNN, Laplace smoothing parameter (alpha) in NB, maximum depth of the tree in DT, number of estimators in RF, learning rate in AdaBoost and GBDT, minimum sum of instance weight of the child tree in XGBoost and the size of hidden layer in MLP.

A high accuracy in training dataset normally indicates a desirable optimization effect while a high validation or test accuracy represents the trained model has a good ability to generalize. Meanwhile, if the training accuracy and validation accuracy vary too much, the model might be overfitting (Xing et al., 2020). Figs. 2–11 presented the hyperparameter tuning process in this study, and it could be observed for dichotomous problem, there is a huge difference in the performances of these ten models for classifying different crash severity dataset.

For QDA model, with the increase of regularization parameter, the accuracy presented an upward tendency, while different crash severity sub-datasets are related to different best regularization parameters. Through the fine-tuning of regularization parameters, there is a huge improvement in the performance of QDA model. However, in SVM model, the accuracy is not sensitive to the change of gamma, which defines how much influence a single training sample has. Likewise, the accuracy curves in MLP models also display a relatively stable trend. With regard to KNN, it was found that as the number of nearest neighbors increases, the training accuracy tends to present a downward, whereas the validation and test accuracy shows an upward before reaching a threshold. The accuracy curves of NB model in SEV1 dataset demonstrated a fluctuant trend with the highest value of validation accuracy obtained at alpha = 140. As for tree-based models, including DT, RF, AdaBoost, GBDT, XGBoost, it was found that deeper trees might not contribute to better classification performance. And for the ensembling tree-based models like AdaBoost and GBDT, the accuracy of models is related to the learning rate, and when the learning rate

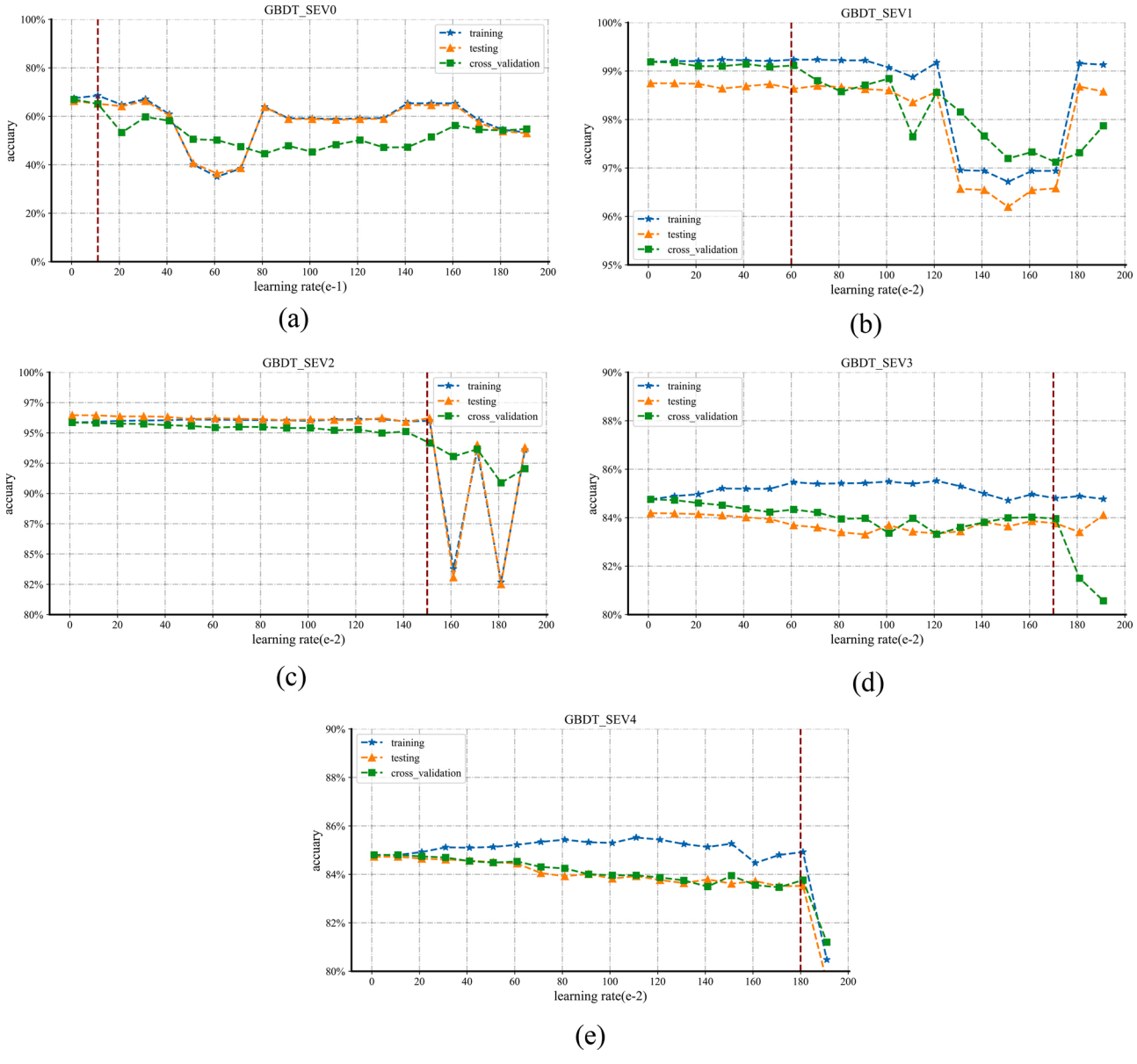


Fig. 5. Hyperparamter tuning in GBDT:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

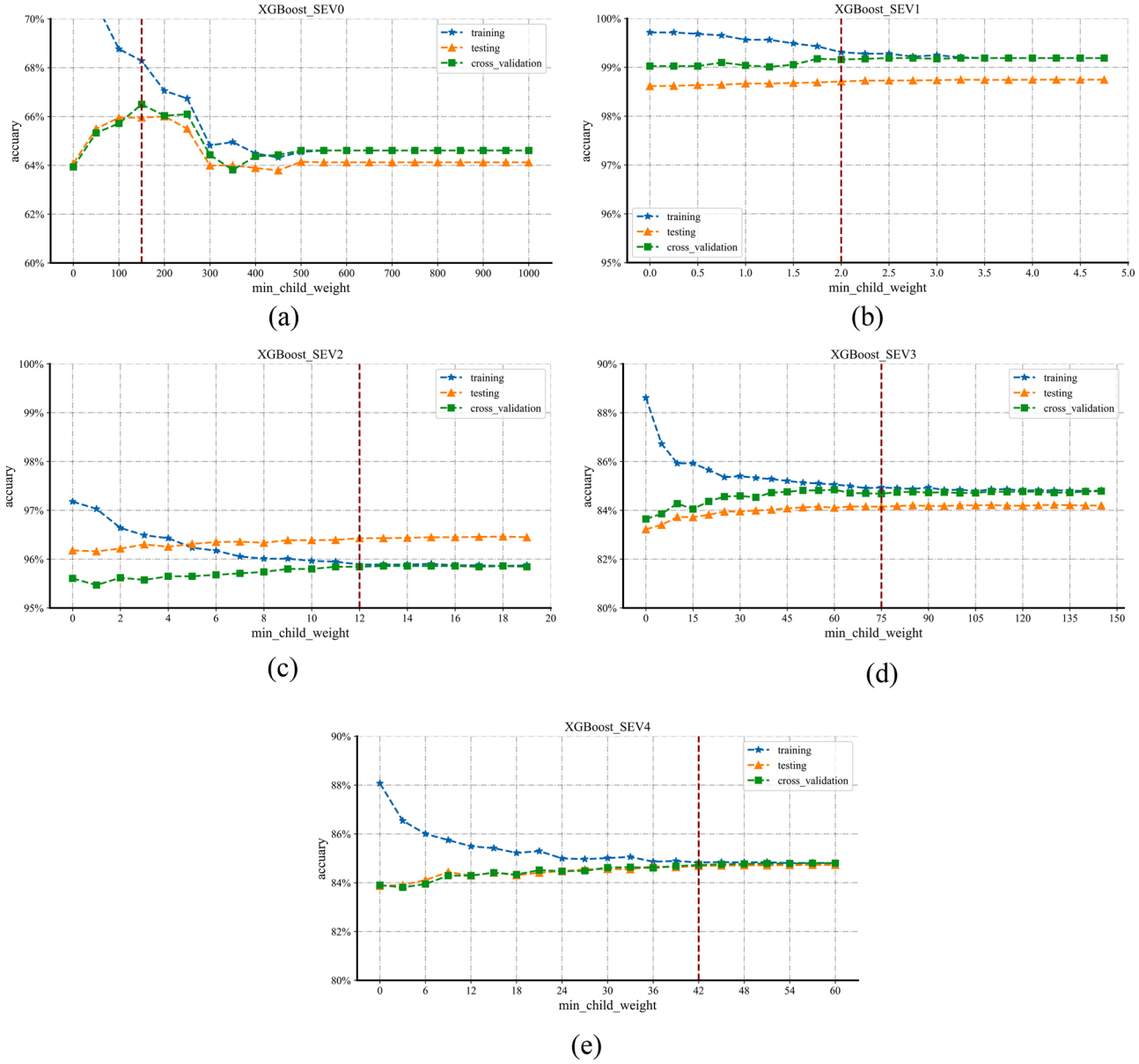


Fig. 6. Hyperparamter tuning in XGBoost:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

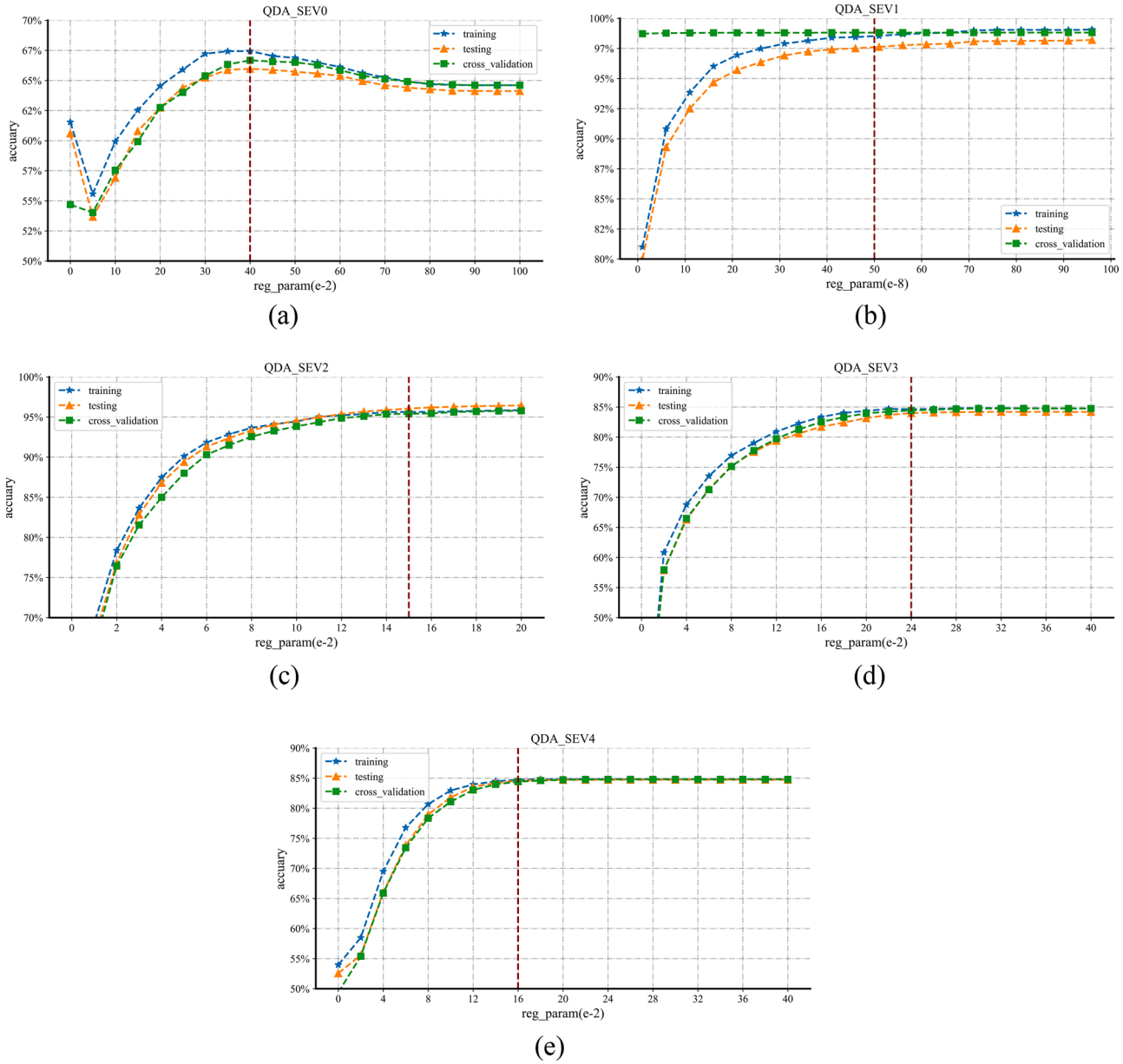


Fig. 7. Hyperparamter tuning in QDA:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

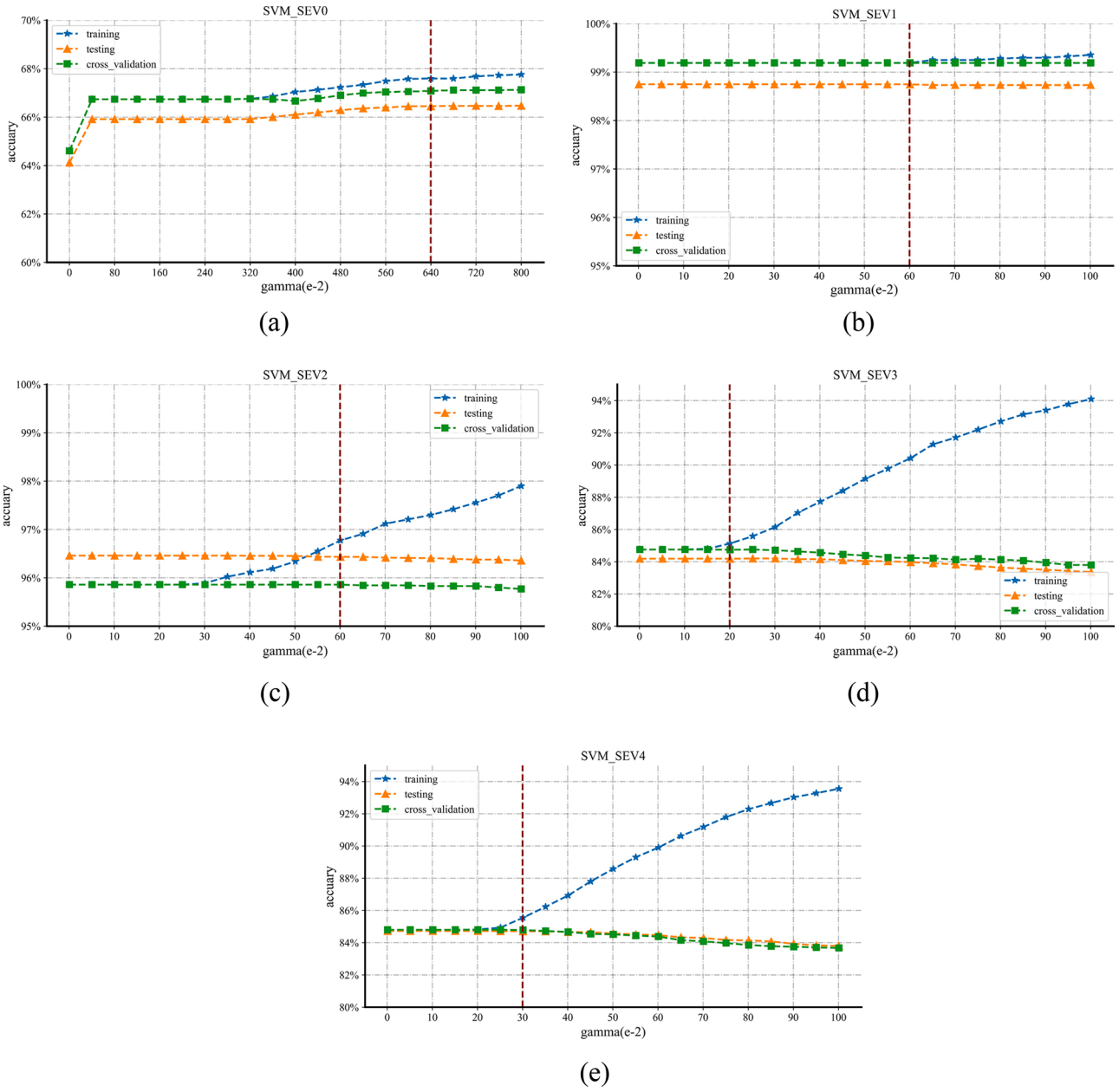


Fig. 8. Hyperparamter tuning in SVM:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

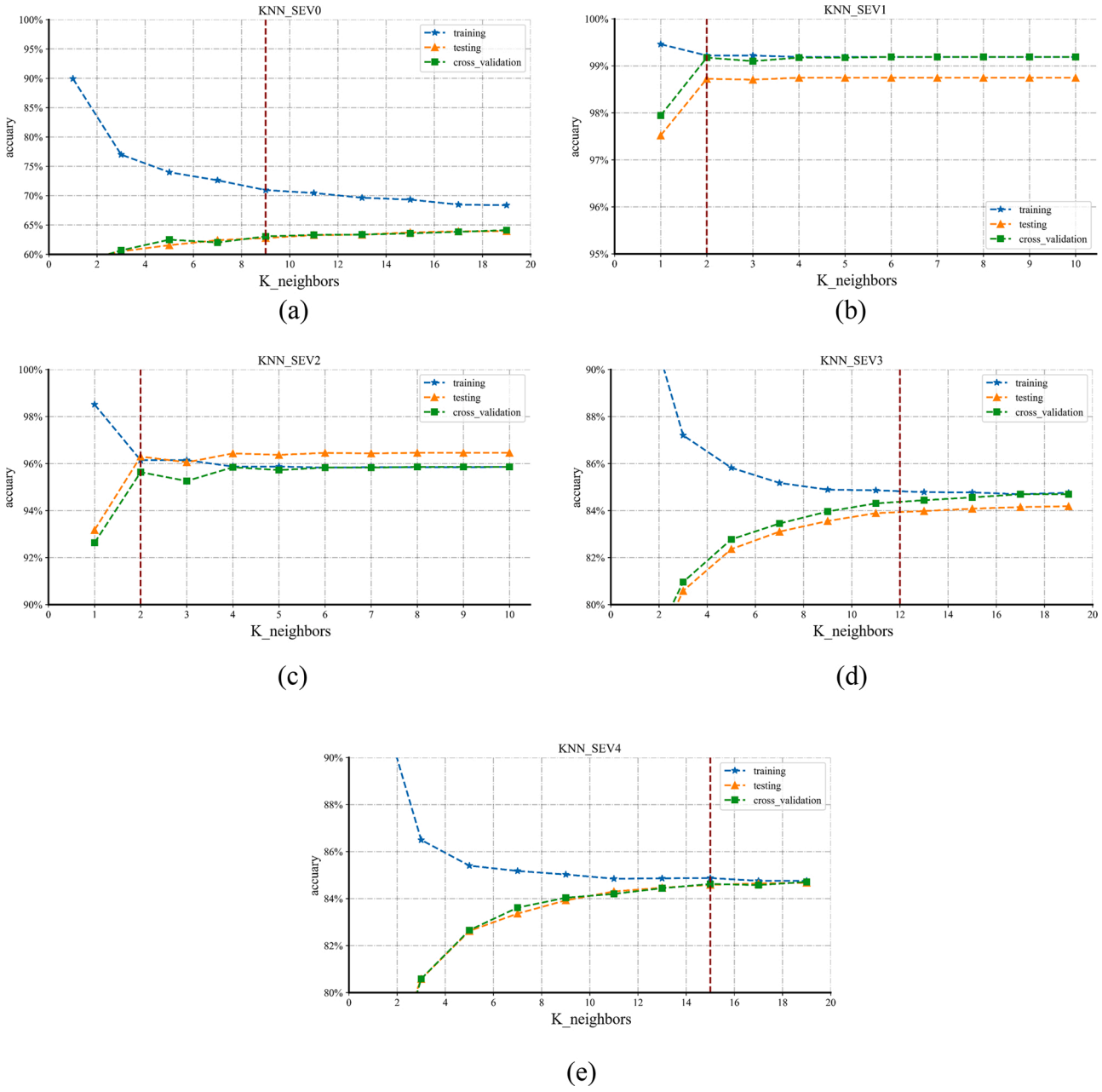


Fig. 9. Hyperparamter tuning in KNN:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

exceeds a certain threshold, the performance of model will decline sharply, since a bigger learning rate might lead to non-convergent in the optimizing process of the model.

4.2. Results of model performances

Based on the result displayed in Figs.2–11 with the GridSearch method, the final performances with best hyperparameters of each model were determined with a high accuracy of training process under the premise of a good validation accuracy (Xing et al., 2020), as shown in Tables 2–6.

As shown in Table 2, these ten models all performed well in fatal crash prediction, no matter in training, validation or test datasets, with the average prediction accuracy reaching 99.27 %, 99.04 % and 98.66 %, respectively. As for severe crash prediction accuracy, its average values of training, validation and test datasets are 96.40 %, 95.52 %, 96.14 %, respectively, marginally lower than those obtained in fatal crash dataset.

From Tables 4–6, it was found that the performances of ten models in accurately predicting less severe crashes are not that satisfactory, while for PDO crash dataset, worse results were obtained. The differences of prediction performances among different crash severities might be attributable to the unbalanced sample distribution. Besides, it's noteworthy that the prediction accuracy distribution based on different crash severity dataset is totally contradictory to the previous study conducted by Tang et al. (2019), where less severe crash dataset obtained higher prediction accuracy. More research could be implemented to further investigate the intrinsic reasons that lead to this difference.

Comparing the performances of different models from a comprehensive perspective in this study, it could be considered that ML methods are marvelously capable of predicting fatal and severe crashes, with the prediction accuracy beyond 95 %, even reaching 99 %, which can serve as a powerful instrument for accident warning and prevention.

What concerns to the tree-based models versus non-tree-based

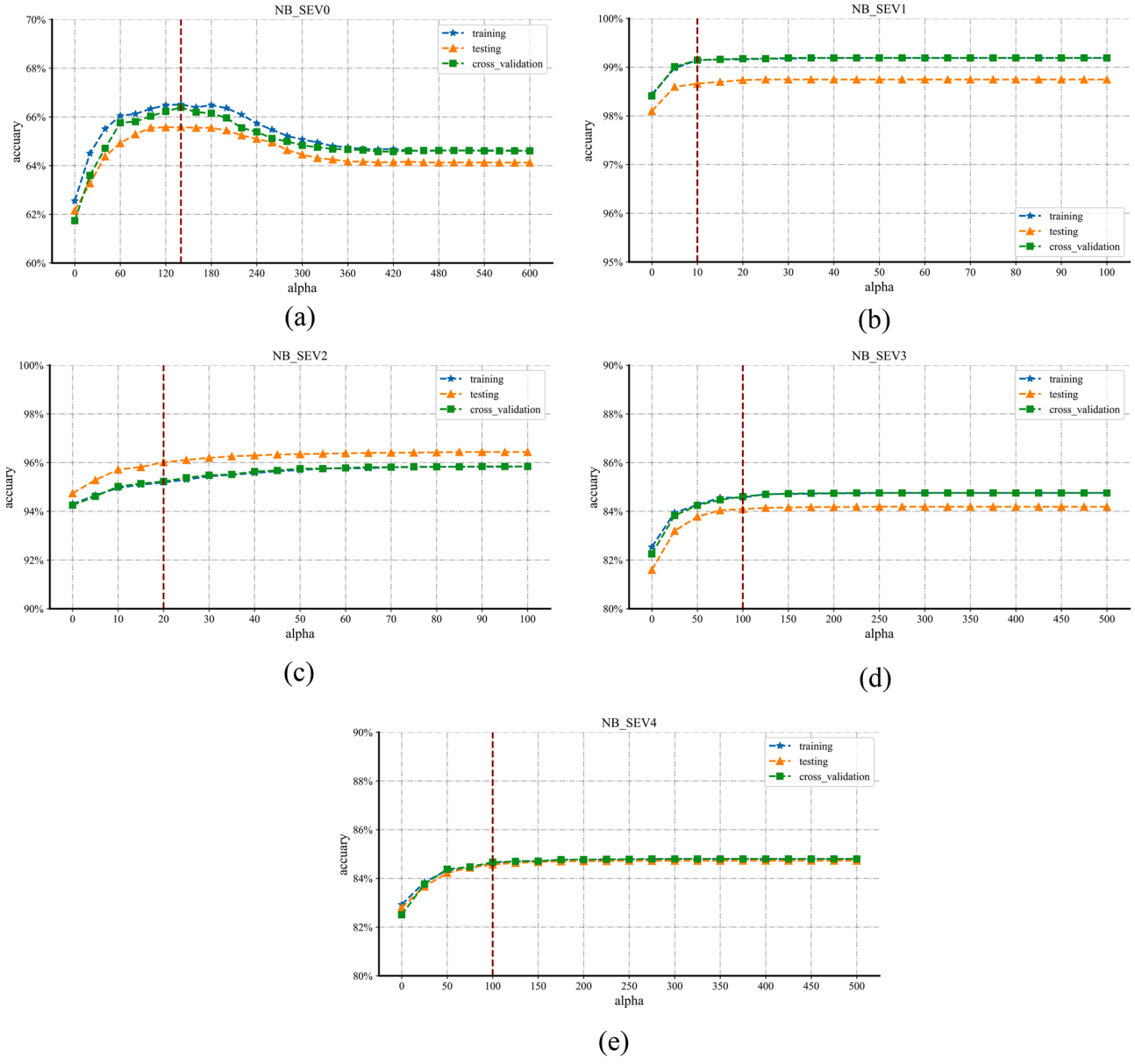


Fig. 10. Hyperparamter tuning in NB:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

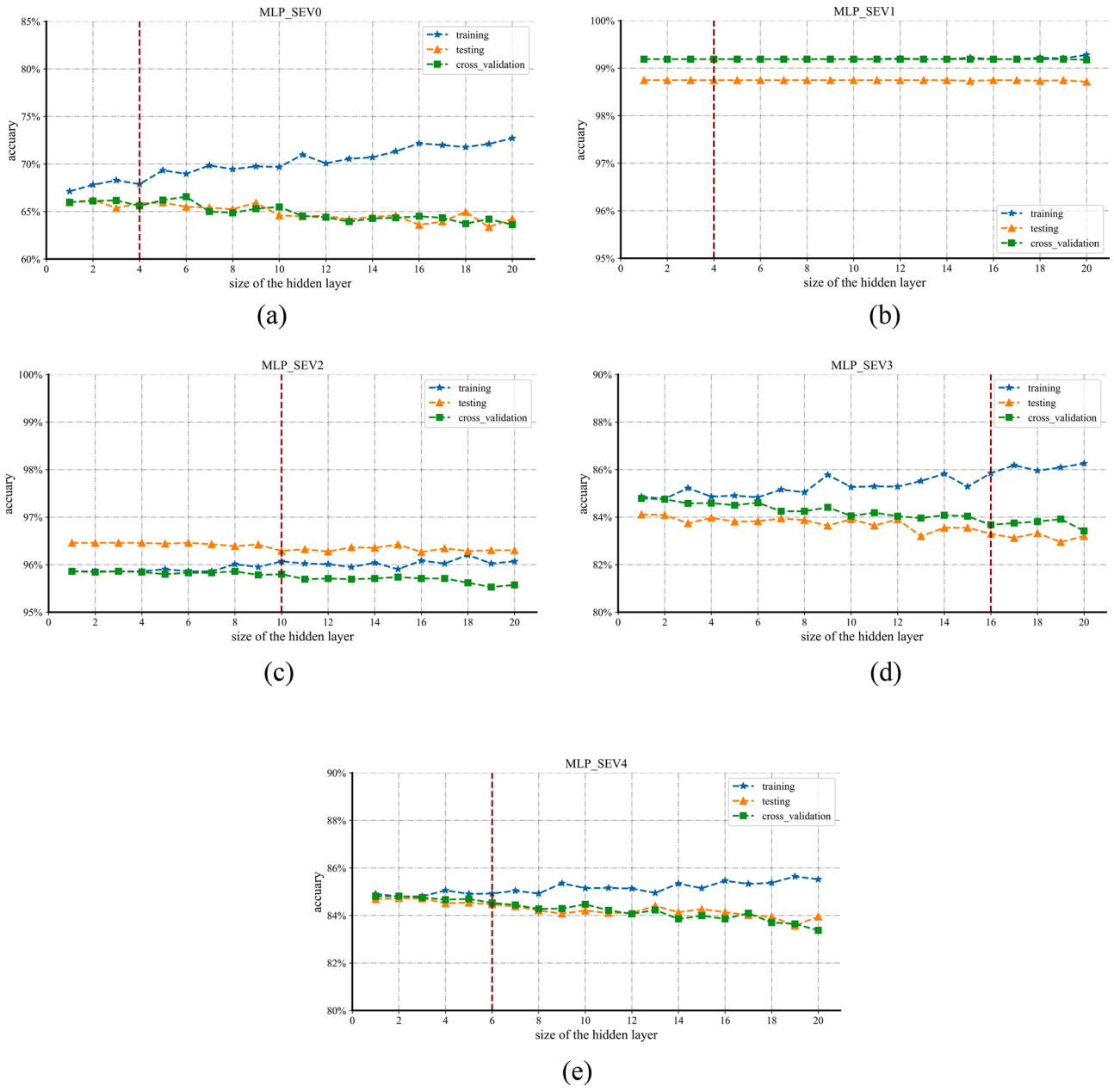


Fig. 11. Hyperparamter tuning in MLP:(a)-(e) SEV0,SEV1,SEV2,SEV3,SEV4.

models, it can be observed that RF model obtained the highest prediction accuracy in training set of all crash precidtion models, not much different from MLP. However, it seems that the generalization ability of RF is not as good as other tree-based models, which is consistent with some previous studies (e.g., Ding et al., 2017; Tang et al., 2019). For SEV1 dataset, the average training accuracy of tree-based models was 99.37 %, while that of non-tree-based models was 99.11 %. Tree-based models also obained higher accuracy in cross-validation and test datasets, on average of 99.11 % and 98.71 %, compared to 98.98 % and 98.61 % of non-tree-based models. Note that this phenomenon can also can be observed in other datasets including SEV0, SEV2, SEV3 and SEV4, indicating that tree-based models might be more appropriate for accurately predicting the occuence and severity of crashes. Note that in SEV0 dataset, nearly all models performed less satisfactory except DT and MLP in the training set, presenting a superior optimazation result, while in test dataset, the models failed to obtain a good generalization result.

4.3. Results of feature importance extraction

In order to identify the contribution of different crash-related features to the prediction results of different crash severities and understand the inherent relationship between various features and crash severities, feature importance extraction based on RF model was undertaken in this paper, as the result of model comparison and selection. The results of feature importance analysis based on five crash severity datasets are presented in Fig.12. On the whole, feature importance distribution varies in each crash severity dataset.

For fatal crash dataset, as shown in Fig.12**(a)**, it could be obviously observed that *urban freeways, beyond shoulder (both right and left), Monday* have a relatively high contribution for the prediction results, indicating under these conditions, the possibility of the occurrence of fatal crashes might increase, which gives the indication that more countermeasures targeted at these factors should be developed to reduce fatal crashes. With respect to severe crash dataset presented in Fig.12 **(b)**, more features are involved with the value of feature importance

Table 2
Model accuracy of SEV1 dataset.

SEV1	Training	Cross_validation	Test
Tree-based models			
DT	99.18%	99.18%	98.75%
RF	99.71%	99.01%	98.65%
AdaBoost	99.19%	99.19%	98.74%
GBDT	99.20%	99.11%	98.74%
XGBoost	99.56%	99.04%	98.67%
Non-tree-based models			
QDA	99.05%	98.84%	98.49%
SVM	99.19%	99.19%	98.75%
KNN	99.19%	99.17%	98.75%
NB	98.71%	98.72%	98.43%
MLP	99.70%	98.97%	98.62%
Average	99.27%	99.04%	98.66%

Note: deeper color indicates higher score.

Table 3
Model performance of SEV2 dataset.

SEV2	Training	Cross_validation	Test
Tree-based models			
DT	95.87%	95.86%	96.49%
RF	98.93%	95.44%	96.02%
AdaBoost	95.86%	95.84%	96.46%
GBDT	95.98%	95.80%	96.44%
XGBoost	97.03%	95.47%	96.16%
Non-tree-based models			
QDA	95.86%	95.86%	96.46%
SVM	96.32%	95.86%	96.45%
KNN	95.87%	95.72%	96.37%
NB	94.31%	94.31%	94.86%
MLP	97.94%	95.02%	95.67%
Average	96.40%	95.52%	96.14%

Note: deeper color indicates higher score.

Table 4
Model performance of SEV3 dataset.

SEV3	Training	Cross_validation	Test
Tree-based models			
DT	84.76%	84.65%	84.19%
RF	95.36%	82.61%	82.27%
AdaBoost	84.74%	84.66%	84.03%
GBDT	85.13%	84.62%	84.12%
XGBoost	88.13%	83.45%	83.11%
Non-tree-based models			
QDA	84.75%	84.75%	84.18%
SVM	89.14%	84.38%	84.04%
KNN	84.83%	84.59%	84.08%
NB	82.56%	82.20%	81.64%
MLP	90.43%	80.99%	81.50%
Average	86.98%	83.69%	83.32%

Note: deeper color indicates higher score.

over than 0 and thus the peak value in feature importance of this dataset is lower than that of fatal one. Relatively, *Friday, Saturday, Monday, male driver, rural freeways, urban freeways* have an important impact on the occurrence of severe crashes. Note that *Monday* and *urban freeways* were found to be significant in both fatal and severe crashes, which deserves more attention. To apply this finding in practice, some warning slogans could be added to dynamic message sign (DMS) in urban freeways on Monday, given that people are often of post weekend syndrome and might be more tired and careless on Monday, which could potentially lead to more severe crashes. A recent study also emphasized the importance of identifying the generation mechanism of severe urban crashes (Intini et al., 2020).

Regarding less severe injury datasets (i.e., SEV3 and SEV4), *urban freeways* also display a high correlation with the prediction results.

Table 5
Model performance of SEV4 dataset.

SEV4	Training	Cross_validation	Test
Tree-based models			
DT	84.82%	84.80%	84.73%
RF	95.11%	82.59%	82.95%
AdaBoost	84.81%	84.77%	84.74%
GBDT	85.03%	84.65%	84.67%
XGBoost	87.47%	83.80%	83.77%
Non-tree-based models			
QDA	84.80%	84.80%	84.72%
SVM	88.58%	84.51%	84.57%
KNN	84.87%	84.62%	84.58%
NB	82.93%	82.67%	82.90%
MLP	89.98%	81.14%	81.80%
Average	86.84%	83.84%	83.94%

Note: deeper color indicates higher score.

Table 6
Model performance of SEV0 dataset.

SEV0	Training	Cross_validation	Test
Tree-based models			
DT	67.29%	67.23%	66.27%
RF	91.46%	62.82%	63.21%
AdaBoost	66.83%	66.30%	66.01%
GBDT	68.21%	66.69%	66.4%
XGBoost	78.24%	63.44%	64.21%
Non-tree-based models			
QDA	66.87%	66.51%	65.72%
SVM	67.98%	67.07%	66.52%
KNN	67.05%	64.24%	64.49%
NB	62.64%	61.92%	62.26%
MLP	81.00%	62.07%	62.21%
Average	71.76%	64.83%	64.73%

Note: deeper color indicates higher score.

Similar to critical factors related to fatal crashes, SEV3 crashes are obviously influenced by roadway characteristics including *beyond shoulders (both right and left)*. Additionally, it seems that *peak* is also associated with the occurrence of less severe injury crashes, and in SEV4 dataset, *daytime* plays a relatively critical role. In addition to *Friday*, *Saturday* and *Sunday*, *Tuesday* appears to be determinant in PDO crashes. Also, the *overturned* feature display a relatively high correlation to PDO crashes relative to other types of crashes, while in previous studies, vehicle overturn was considered to be highly related with more severe crashes (e.g., Fountas et al., 2018; Wu et al., 2014; Xie et al., 2012). It might be due to that the dataset applied in this paper is spatiotemporally different from these past work, and the contributing factors of crashes are of spatiotemporal instability (Behnood and Mannering, 2015; Dabbour et al., 2020, 2017).

5. Conclusion

Crashes are considered as a global matter that triggers fatalities and serious injuries (Alkheder et al., 2020), especially for those single-vehicle-related ones (Gong and Fan, 2017; Li et al., 2019a; Wu et al., 2014, 2016; Xie et al., 2012). This paper utilized both tree-based and other non-parameter models to predict crash severities related to sophisticated contributing factors and investigated whether there is a difference in models' ability to correctly identify crash severity based on different prediction tasks. By dividing five single-vehicle crash dataset according to different crash injury severity into training, validation and test sets, respectively and taking the accuracy as the major evaluation metric, the best hyper-parameters of ten models were selected with the GridSearch method, including QDA, SVM, KNN, NB, DT, RF, AdaBoost, GBDT, XGBoost, MLP. Through a comparison analysis, the main findings can be concluded as follows:

(1) For crash severity prediction in the same severity level, the ten models didn't exhibit a huge performance difference; however, the performances of model did vary among different dataset, with an average training accuracy of 99.27 %, 96.4 %, 86.98 %, 86.84 %, 71.76 % in fatal injury, severe injury, visible injury, complaint of pain, PDO crash datasets, respectively.
(2) RF model outperformed other models in training prediction accuracy for all five sub-datasets, while it didn't have the same accuracy in cross validation and test sets, indicating that RF model might have a lower generalization ability.
(3) Tremendous previous studies have been conducted to predict crash severity with tree-based models including DT, RF, AdaBoost, GBDT, MLP; however, this paper, proved that other ML models are also appropriate for crash severity analysis to consider it as a dichotomous problem; especially for those more severe accidents, all of models proposed in this study display a relatively high accuracy.

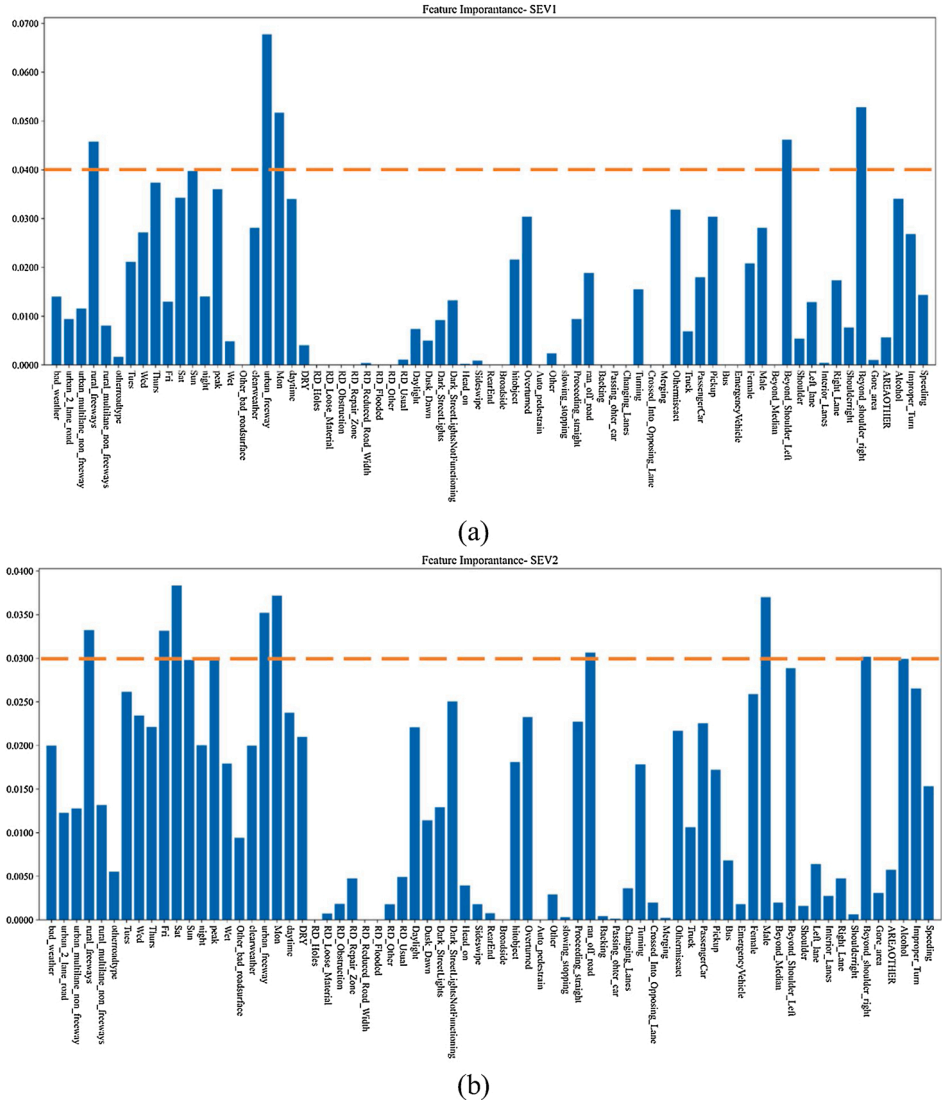
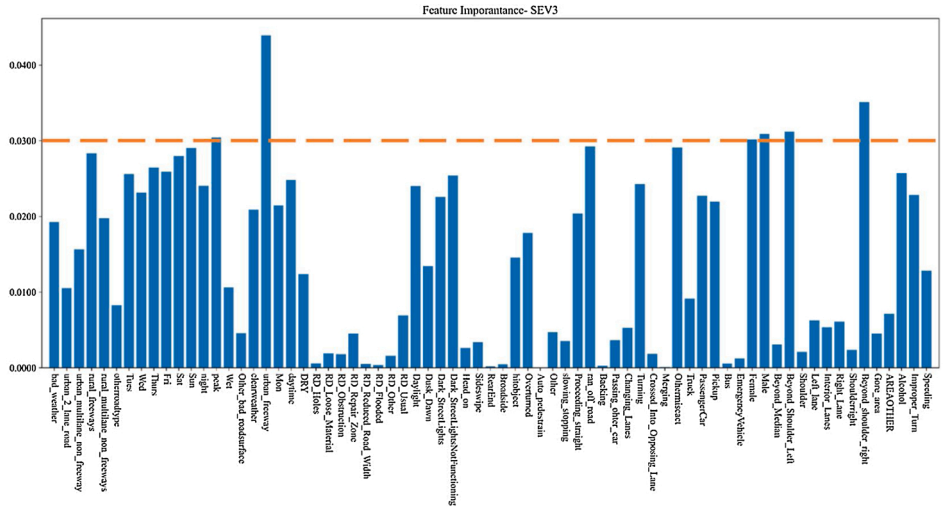
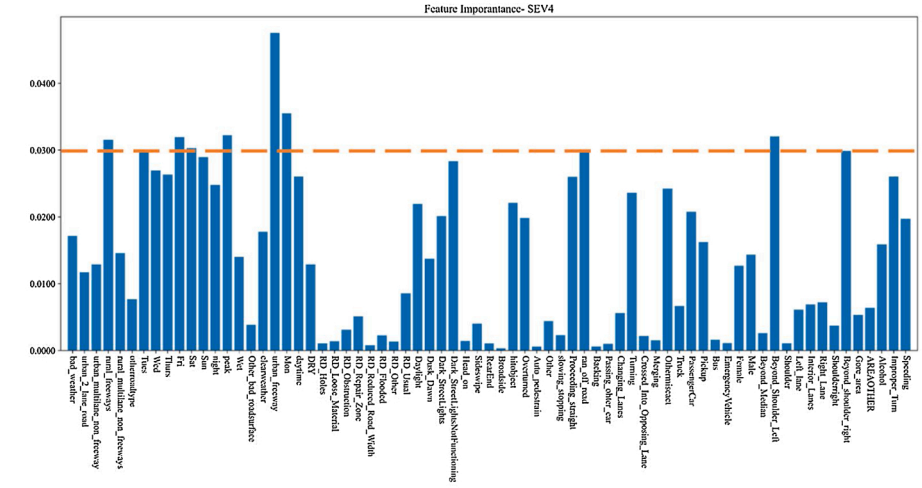


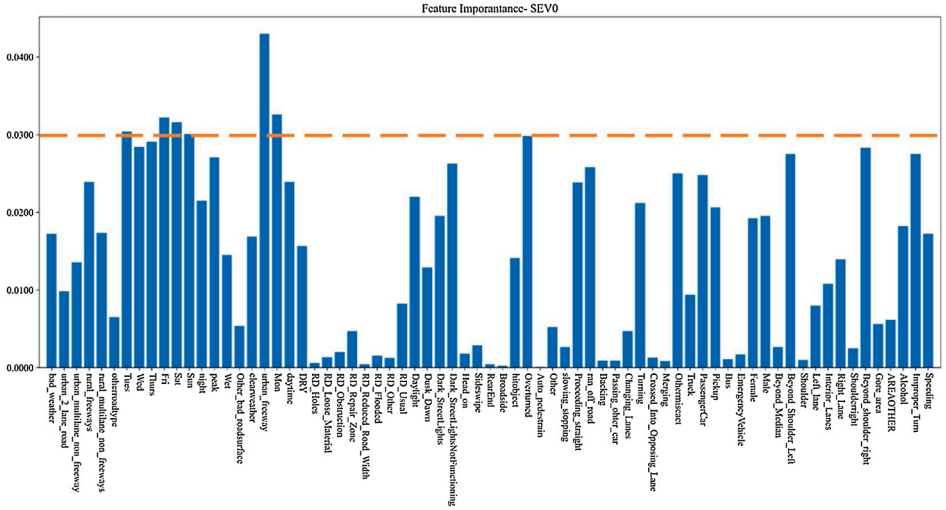
Fig. 12. Crash feature importance based on different severity scales: (a) fatal injury; (b) severe injury; (c) visible injury; (d) complaint of pain; and (e) PDO.



(c)



(d)



(e)

Fig. 12. (*continued*).

Besides, to further identify the intrinsic contributing factors that determine the crash severity, feature importance was extracted in this study, with the following findings:

(1) For all crash severities, *urban freeways* were significant in the occurrence of crashes, indicating that additional attention should be given to urban freeway accident prevention.

(2) Besides *urban freeways,* the crashes related to *rural freeways* and *beyond shoulders* were more severe compared to those with other roadway characteristics. Regarding time characteristics, weekend is of higher risk for single-vehicle crashes, including *Friday, Saturday, Sunday.*

The contributions of this paper can be summarized as two bullets: (1) to provide useful information for model selection and tuning in accident severity prediction; and (2) to facilitate the development of the policies and countermeasures to reduce the accident severity.

However, it should be noted that this paper is not free of limitations, more research could be implemented from these two aspects: (1) to further consider the influences that temporal instability of contributing features has on the model performances with more input sample distributed in different years; and (2) to explore the availability of deep learning (DL) models such as CNN, RNN, LSTM in crash severity prediction and compare their performance difference with traditional ML methods.

CRediT authorship contribution statement

Xintong Yan: Methodology, Software, Validation, Formal analysis, Writing - original draft. **Jie He:** Conceptualization, Supervision, Project administration, Funding acquisition. **Changjian Zhang:** Investigation, Writing - original draft. **Ziyang Liu:** Investigation, Resources. **Boshuai Qiao:** Data curation. **Hao Zhang:** Investigation.

Declaration of Competing Interest

The authors reported no declarations of interest.

Acknowledgments

The authors would like to thank National Natural Science Foundation of China (Grant No. 52072069 and 51778141), Transportation Department of Henan Province (Grant No. 2018G7), and Jiangsu Creative PHD student sponsored project (KYCX20_0138). Their assistance is gratefully acknowledged.

References

Abellán, J., López, G., Oña, J.De., 2013. Analysis of traffic accident severity using Decision Rules via Decision Trees. Expert Syst. Appl. 40, 6047–6054. https://doi.org/10.1016/j.eswa.2013.05.027.

Alkheder, S., Taamneh, M., Taamneh, S., 2017. Severity prediction of traffic accident using an artificial neural network. J. Forecast. 36 (1), 100–108. https://doi.org/10.1002/for.2425.

Alkheder, Sharaf, Alrukaibi, F., Aiash, A., 2020. Risk analysis of traffic accidents' severities : An application of three data mining models. ISA Trans. https://doi.org/10.1016/j.isatra.2020.06.018 (xxxx).

Arnold, M.G., Morgan, A., 2019. One-Hot residue logarithmic number systems. 2019 29th International Symposium on Power and Timing Modeling, Optimization and Simulation (PATMOS) 97–102. https://doi.org/10.1109/PATMOS.2019.8862159.

Behnood, Ali, Mannering, F.L., 2015. Analytic Methods in Accident Research The temporal stability of factors affecting driver-injury severities in single-vehicle crashes : Some empirical evidence. Anal. Methods Accid. Res. 8, 7–32. https://doi.org/10.1016/j.amar.2015.08.001.

Behnood, A., Mannering, F.L., 2017. The effects of drug and alcohol consumption on driver injury severities in single-vehicle crashes. Traffic Inj. Prev. 18 (5), 456–462. https://doi.org/10.1080/15389588.2016.1262540.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breiman, L., Friedman, J., J.Stone, C., Olshen, R.A., 1984. Classification and Regression Trees. Chapman and Hall.

Chang, L., Chien, J., 2013. Analysis of driver injury severity in truck-involved accidents using a non-parametric classification tree model. Saf. Sci. 51 (1), 17–22. https://doi.org/10.1016/j.ssci.2012.06.017.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16 (1), 321–357.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785.

Chen, C., Zhang, G., Cathy, X., Ci, Y., Huang, H., Ma, J., et al., 2016. Driver injury severity outcome analysis in rural interstate highway crashes : a two-level Bayesian logistic regression interpretation. Accid. Anal. Prev. 97, 69–78. https://doi.org/10.1016/j.aap.2016.07.031.

Chollet, F., 2018. Deep Learning With Python. Manning Publications., New York.

Dabbour, E., Easa, S., Haider, M., 2017. Using fixed-parameter and random-parameter ordered regression models to identify significant factors that affect the severity of drivers' injuries in vehicle-train collisions. Accid. Anal. Prev. 107 (December 2016), 20–30.

Dabbour, E., Dabbour, O., Martinez, A.A., 2020. Temporal stability of the factors related to the severity of drivers' injuries in rear-end collisions. Accid. Anal. Prev. 142 (October 2019).

Das, A., Abdel-aty, M., Pande, A., 2009. Using conditional inference forests to identify the factors affecting crash severity on arterial corridors. J. Safety Res. 40 (4), 317–327. https://doi.org/10.1016/j.jsr.2009.05.003.

Ding, C., Luan, S., Wang, Y., Wang, Y., 2017. Prioritizing influential factors for freeway incident clearance time prediction using the gradient boosting decision trees method. Ieee Trans. Intell. Transp. Syst. 18 (9), 2303–2310.

Federal Highway Administration(FHWA), 2020. Highway Safety Information System (HSIS) Database. Retrieved August 29, 2020, from. http://www.hsisinfo.org/.

Feng, S., Li, Z., Ci, Y., Zhang, G., 2016. Risk factors affecting fatal bus accident severity : Their impact on different types of bus drivers. Accid. Anal. Prev. 86, 29–39. https://doi.org/10.1016/j.aap.2015.09.025.

Figueira, C., Pitombo, C.S., Tadeu, P., Oliveira, S.De, Paula, A., Larocca, C., 2017. Case Studies on Transport Policy Identi fi cation of rules induced through decision tree algorithm for detection of traf fi c accidents with victims : A study case from Brazil. Case Stud. Transp. Policy 5 (2), 200–207. https://doi.org/10.1016/j.cstp.2017.02.004.

Fountas, G., Anastasopoulos, P.C., Abdel-aty, M., 2018. Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. Anal. Methods Accid. Res. 18, 57–68.

Fountas, G., Fonzone, A., Gharavi, N., Rye, T., 2020. The joint effect of weather and lighting conditions on injury severities of single-vehicle accidents. Anal. Methods Accid. Res. 27.

Freund, Y., Schapire, R.E., 1995. A decision-theoretic generalization of on-line learning and an application to boosting. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 904 (1), 23–37. https://doi.org/10.1007/3-540-59119-2_166.

Gong, L., Fan, W.D., 2017. Modeling single-vehicle run-off-road crash severity in rural areas : Accounting for unobserved heterogeneity and age difference. Accid. Anal. Prev. 101, 124–134. https://doi.org/10.1016/j.aap.2017.02.014.

Haque, M., Chin, H.C., Debnath, A.K., 2012. An investigation on multi-vehicle motorcycle crashes using log-linear models. Saf. Sci. 50 (2), 352–362. https://doi.org/10.1016/j.ssci.2011.09.015.

Hongpu, F., Beiji, Z., 2020. A fast training method for AdaBoost classifier. Journal of Yunnan University 42 (1), 50–57.

Intini, P., Berloco, N., Fonzone, A., Fountas, G., Ranieri, V., 2020. The influence of traffic, geometric and context variables on urban crash types : A grouped random parameter multinomial logit approach. Anal. Methods Accid. Res. 28, 100141 https://doi.org/10.1016/j.amar.2020.100141.

Iranitalab, A., Khattak, A., 2017. Comparison of four statistical and machine learning methods for crash severity prediction. Accid. Anal. Prev. 108 (February), 27–36.

Jung, S., Jang, K., Yoon, Y., Kang, S., 2014. Contributing factors to vehicle to vehicle crash frequency and severity under rainfall. J. Safety Res. 50, 1–10.

Kashani, A.T., Rabieyan, R., Besharati, M.M., 2014. A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers. J. Safety Res. 51, 93–98. https://doi.org/10.1016/j.jsr.2014.09.004.

Kim, J., Ulfarsson, G.F., Kim, S., Shankar, V.N., 2013. Driver-injury severity in single-vehicle crashes in California : A mixed logit analysis of heterogeneity due to age and gender. Accid. Anal. Prev. 50, 1073–1081.

Li, Z., Chen, C., Wu, Q., Zhang, G., Liu, C., Prevedouros, P.D., Ma, D.T., 2018. Analytic Methods in Accident Research Exploring driver injury severity patterns and causes in low visibility related single-vehicle crashes using a finite mixture random parameters model. Anal. Methods Accid. Res. 20 (August), 1–14. https://doi.org/10.1016/j.amar.2018.08.001.

Li, L., Prato, C.G., Wang, Y., 2020. Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: a sequential approach of multivariate imputation by chained equations and random forest classifier. Accid. Anal. Prev. 146 https://doi.org/10.1016/j.aap.2020.105744.

Li, Z., Ci, Y., Chen, C., Zhang, G., Wu, Q., Qian (Sean), Z., et al., 2019a. Investigation of driver injury severities in rural single-vehicle crashes under rain conditions using mixed logit and latent class models. Accid. Anal. Prev. 124, 219–229. https://doi.org/10.1016/j.aap.2018.12.020.

Li, Z., Wu, Q., Ci, Y., Chen, C., Chen, X., Zhang, G., 2019b. Using latent class analysis and mixed logit model to explore risk factors on driver injury severity in single-vehicle crashes. Accid. Anal. Prev. 129 (April), 230–240.

Mahmodi, K., Mostafaei, M., Mirzaee-ghaleh, E., 2019. Detection and classification of diesel-biodiesel blends by LDA, QDA and SVM approaches using an electronic nose. Fuel 258 (September), 116114. https://doi.org/10.1016/j.fuel.2019.116114.

Martensen, H., Dupont, E., 2013. Comparing single vehicle and multivehicle fatal road crashes : A joint analysis of road conditions, time variables and driver characteristics. Accid. Anal. Prev. 60, 466–471. https://doi.org/10.1016/j.aap.2013.03.005.

Menaka, D., Suresh, L.P., Kumar, S.S.P., 2014. Land cover classification of multispectral satellite images using QDA classifier. 2014 International Conference on Control,

Instrumentation, Communication and Computational Technologies (ICCICCT) 1383–1386.
Muhammad, F., Rashid, N., Akhtar, H., Muhammad, Z., Gilani, S.O., Ansari, U., Subjects, A., 2014. Evaluation of LDA, QDA and decision trees for multifunctional controlled below elbow prosthetic limb using EMG signals. 2014 International Conference on Robotics and Emerging Allied Technologies in Engineering (ICREATE), Robotics and Emerging Allied Technologies in Engineering (ICREATE), 2014 International Conference On, 115–117. https://doi.org/10.1109/iCREATE.2014.6828350.
Naik, B., Tung, L., Zhao, S., Khattak, J., 2016. Weather impacts on single-vehicle truck crash injury severity. J. Safety Res. 58, 57–65. https://doi.org/10.1016/j.jsr.2016.06.005.
National Highway Traffic Safety Administration, 2019. Traffic Safety Facts Annual Report Tables. Retrieved from. https://cdan.nhtsa.gov/SASStoredProcess/guest.
Osman, M., Mishra, S., Paleti, R., 2018. Injury severity analysis of commercially-licensed drivers in single-vehicle crashes : Accounting for unobserved heterogeneity and age group differences. Accid. Anal. Prev. 118 (May), 289–300. https://doi.org/10.1016/j.aap.2018.05.004.
Shi, Q., Abdel-aty, M., 2015. Big Data applications in real-time traffic operation and safety monitoring and improvement on urban expressways. Transp. Res. Part C Emerg. Technol. 58, 380–394. https://doi.org/10.1016/j.trc.2015.02.022.
Taleshmekaeil, D.K., Safari, A., Kong, Y., 2012. Using one hot residue number system (OHRNS) for digital image processing. The 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP 2012) 64–67. https://doi.org/10.1109/AISP.2012.6313719.
Tang, J., Liang, J., Han, C., Li, Z., 2019. Crash injury severity analysis using a two-layer Stacking framework. Accid. Anal. Prev. 122 (October 2018), 226–238. https://doi.org/10.1016/j.aap.2018.10.016.
Tang, J., Zheng, L., Han, C., Yin, W., Zhang, Y., Zou, Y., 2020. Statistical and machine-learning methods for clearance time prediction of road incidents : A methodology review. Anal. Methods Accid. Res. 27.
Wu, J., Abdel-aty, M., Yu, R., Gao, Z., 2013. A novel visible network approach for freeway crash analysis. Transp. Res. Part C Emerg. Technol. 36, 72–82. https://doi.org/10.1016/j.trc.2013.08.005.
Wu, Q., Chen, F., Zhang, G., Cathy, X., Wang, H., Bogus, S.M., 2014. Mixed logit model-based driver injury severity investigations in single- and multi-vehicle crashes on rural two-lane highways. Accid. Anal. Prev. 72, 105–115.
Wu, Q., Zhang, G., Zhu, X., Cathy, X., Tarefder, R., 2016. Analysis of driver injury severity in single-vehicle crashes on rural and urban roadways. Accid. Anal. Prev. 94, 35–45.
Xie, Y., Zhao, K., Huynh, N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accid. Anal. Prev. 47, 36–44. https://doi.org/10.1016/j.aap.2011.12.012.
Xing, L., He, J., Li, Y., Wu, Y., Yuan, J., Gu, X., 2020. Comparison of different models for evaluating vehicle collision risks at upstream diverging area of toll plaza. Accid. Anal. Prev. 135 (September 2019).
Xu, C., Wang, W., Liu, P., 2013. A genetic programming model for real-time crash prediction on freeways. IEEE Transactions on Intelligent Transportation Systems, Intelligent Transportation Systems, IEEE Transactions on, IEEE Trans. Intell. Transport. Syst. 14 (2), 574–586. https://doi.org/10.1109/TITS.2012.2226240.
Yu, B., Wang, Y.T., Yao, J.B., Wang, J.Y., 2016. A comparison of the performance of ANN and SVM for the prediction of traffic accident duration. Neural Netw. World 26 (3), 271–288. https://doi.org/10.14311/NNW.2016.26.015.
Yu, H., Li, Z., Zhang, G., Liu, P., 2019. A latent class approach for driver injury severity analysis in highway single vehicle crash considering unobserved heterogeneity and temporal influence. Anal. Methods Accid. Res. 24.
Zhang, H., Su, J., 2008. Naive Bayes for optimal ranking. J. Exp. Theor. Artif. Intell. 20 (2), 79–93. https://doi.org/10.1080/09528130701476391.
Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Accid. Anal. Prev. 43, 49–57. https://doi.org/10.1016/j.aap.2010.07.007.