# Analysing the Road Traffic Accident Severity using Ensemble methods

Pramir K C
*Department of Computer Science*
*Dalhousie University*
Halifax, Canada
pr260461@dal.ca

***Abstract*—Road traffic accidents are the major cause of life-threatening injuries and fatalities all over the world. The aim of this experiment is to understand the risk factors and how they impact the severity of accidents. This is quite vital in formulating policies to mitigate the frequency and severity of accidents. Ensemble methods were proposed for the classification of 'Road traffic Accidents' severity in this article. XGBoost which works on the principle of boosting and Random Forest (RF) which works on the principle of bagging and random selection of features for building trees were applied in this analysis. Both ensemble methods help reduce bias and variance and have found to perform well with large, noisy datasets with many features. Based on the experiments Random Forest and XGBoost performed very similar in terms of overall performance. However, for the classification performance for individual class of traffic accident severity, XGBoost performed better for less frequent classes**

***Index Terms*—Road traffic accident severity, ensemble methods, Random Forest, XGBoost**

## I. Introduction

Road traffic accidents are the major cause of fatalities in the world with 1.3 million deaths every year. In fact, the number one cause of deaths among young people aged 5-29 is due to road accidents [1]. This problem is even worse in low and middle income countries where 90% of all the death casualties come from despite having 54% of total number of vehicles [2].

It comes as a major importance for governments and individuals to identify the risk factors associated with accident severity and determine which risk factors needs to be focused on to mitigate the accident severity.

Some of the major risk factors associated with higher level of traffic accident severity are street conditions, bad visibility, males, speed factor, season of the year, education level, driver fatigue , speed limit and so on. Thus, the risk factors could be broadly classified into human, road , environment and vehicle [2].

Despite the analysis and identification of Road traffic accident severity risk factors by using various Machine Learning (ML) tools, mixed effects for same risk factors have been found depending on which country the data came from or the different paper published from the same country. This makes the task of identifying common risk factors for road traffic accident severity extremely challenging [2].

Thus, the goal of this research is to identify a robust ML tool with high interpretability and explainability in order to classify the accident severity from 'Road Traffic Accidents' data from Kaggle [3]. The model is expected to have high performance metrics along with ability to determine and explain how the most important attributes contribute to the classification of road accident severity for the dataset.

By determining the major risk factors, favourable policies could be implemented so that the risk of high accident severity could be alleviated.

## II. RELATED WORK

### A. *Rough set theory and SVM*

To assess the severity of road traffic accident data from 2006 to 2013 in China, a classification and recognition model was used [4]. 53 attributes were simplified and classified into 5 different broad attributes: human, vehicle, road, environment, and accident. The model used in this research was rough set theory and SVM [4]. In this research, first the rough set theory was used to determine the feature importance based on 5 different kinds of attributes. It was acknowledged that there were several statistical components to large-scale accident data, and that having too many dimensions and levels may impair classification accuracy and computing performance. Rough set theory has advantage as it could handle ambiguous, conflicting, or incomplete data and it doesn't require any prior knowledge and has a strong mathematical foundation [4]. Similarly, SVM was selected as a classification model because according to the Jianfeng et al.[4] it finds the optimal balance between model complexity and learning potential when there is limited sample size. Moreover, they claim that classification speed is quite high for SVM and has high good classification outcome for small dataset. The dataset size for this experiment was 4320 and 53 attributes. By calculating the feature importance from the rough set theory, the Jianfeng et al.[4] were able to reduce the complexity of data by selecting only 30 attributes. The justification for better performance of combined SVM model along with Rough Set theory was the increase in accuracy to 93% from 84% compared to when SVM alone is used. Moreover, the computational speed decreased to almost half from 56 seconds to 30 seconds [4].

For SVM if the parameters C and r (in the case of a Gaussian kernel) are suitably selected, it offers a decent out-of-sample generalisation. This offers robustness to the SVM even when the training sample has some bias [5]. Due to the convex nature of the optimality problem, SVMs offer a special solution. This is a benefit compared to neural networks, which may not be stable across diverse samples since they contain many solutions connected to local minima [5].

The potential problems using SVM would be that the SVM model performs poorly when the margin of separation of target variables is not clear i.e., they are overlapping [6]. Jianfeng et al.[4] simplified the labels into 2 classes: severe accident and general accident. It would be very interesting to find out how SVM used along will Rough set theory will perform when there are multi-level labels for the accident severity.

### *B. Logistic Regression*

Ditcharoen et al.[2] reviewed 16 different published paper to understand the risk factors associated to traffic accident severity along with the machine learning models used for classification of traffic accident severity. It was found that 6 of the papers in the review used logistic regression for classification model. Moreover, it was identified that the speed of the vehicle and human features were the most important variables to determine traffic accident severity. Some other important factors were vehicle type, weather, alcohol consumption and driver's fatigue. However, one important finding was some of the variables have mixed effects depending on the country where the data was obtained. Some explanatory variables had positive or negative effects on the severity of accident depending on if the country where the data was obtained i.e., developing or developed country [2].

The argument for using logistic regression in majority of papers was it is most simple and interpretable classification model with good understanding of parameters and outcome [2]. For example, in logistic regression, we can plot the coefficient of regression and understand how increase/decrease in the coefficient of explanatory variable affects the response variable. Moreover, the benefit of logistic regression is that it examines the connection of all variables simultaneously, avoiding confounding effects [2].

The disadvantage of logistic regression is that it can overfit in high dimensional dataset. Moreover, it does not perform well when there is multicollinearity in the dataset [7]. When two explanatory variables are highly correlated, the impact of both of those variables in the model gets diminished. When large number of variables including less important variables are used, the model either detects erroneous correlations or muddle actual relationships, producing huge standard errors and inaccurate confidence intervals [7].

### *C. Naive Bays, Gradient Boosting trees and Deep Learning*

Raw data were used from Spanish traffic agency over the period of 6 years from 2011 to 2015. Three classification ML tools were used to assess the performance (accuracy, precision and F1 measure for each experiment): Naïve Bayes, Grading boosting trees and Deep Learning [8].

The Naive Bayes is a classification model which uses Bayes theorem to creates a probabilistic model in a quick and easy manner. Cuenca et al.[8] decides to use Naïve Bayes as one of the classification models for predicting traffic accident severity because it is a high-bias, low-variance classifier that can achieve excellent performance with small datasets.

A naive Bayes classifier has the benefit of just requiring a modest quantity of training data to estimate the essential parameters (variable mean and variance). The entire covariance matrix is not required and estimation of only each variable is required as the assumption of independence of variables is made [9].

One of the major shortcomings of Naïve Bayes is that it assumes the independence of the variable while, many of the variables are not independent. This creates a model which treats the redundant, interactive, and noisy variables same as more important features, which leads to the reduction in classification performance [9].

Cuenca et al.[8] decided to use Gradient boosting trees because it is a forward-learning ensemble methods (classification or regression) that rely on progressively more accurate prediction (boosting) to produce projected results. The advantage of boosting method is that it produces hypothesis better than a single random classification model[8].

Cuenca et al.[8] used Convolution Neural Network for the classification task. A simple neural network has three kinds of layers in their architecture. The input layer which processes the data, analyze it and pass it to the hidden layer. The hidden layer that takes input from input layers and hidden layers. Finally output layers where the output is produced from classification [10].

The problem with deep learning model like ANN is that it acts like black box and the mechanistic knowledge of underlying process is hard to assess. Therefore, even if the model reaches great accuracy, it cannot serve as the foundation for further scientific research since it lacks theoretical backing [11].

Based on the experiments by Cuenca et al.[8], the deep learning classification model outperformed other two model. The performance of deep model was not significantly higher compared to Gradient boosting trees after optimization. However, both gradient boosting trees and deep learning classification model outperformed the Naive Bayes model.

## III. Methodology

Based on many methods published in literature, it is clear that simple model like logistic regression have great interpretability with good understanding of parameters and outcome. However, this model can overfit in high dimensional dataset. This could result in low generalization and low performance for testing dataset. Moreover, logistic regression performs badly when there is multicollinerity in dataset [7]. The 'Road Traffic Accidents' had many variables that showed multicollinerity. On the other hand, the deep learning methods like CNN showed better performance. However, the explanability and interpretability of these models were very low for classification of accident severity.

Based on the available solutions from different models, ensemble methods are the best options for 'Road Traffic

Accident' dataset because they reduce the variance and bias of the model. They work on the principle that rather than having one hypothesis that fits the data, having multiple hypothesis created and using those hypotheses to vote for the label of the dataset would provide the best performance. Several empirical experiments have proved that ensemble methods are low variance and low bias model [12]. Ensemble method performs best because it partly solves some of the challenges faced by models that offers single hypothesis. While searching for the space of hypothesis for training data, there could be many hypotheses that provide similar performance on the training data and the learning algorithm should find the best hypothesis. By using the voting method, the risk of choosing a hypothesis that performs badly is highly reduced [12]. In methods like CNN, there is a risk of not finding the best parameters when the model is stuck at the local minima. The ensemble method solves those issues by choosing the weighted combination of local minima [12]. It also solves the representation problem which is not of the existing hypothesis provides the good approximations for the true function f. In this scenario, by taking the weighted votes of all the hypothesis, close approximation of true function f could be made.[12]

Thus, for the 'Road Traffic Accidents' dataset two ensemble methods i.e., Random Forest and XGBoost which works on the principle of Bagging and random selection of Variable and Boosting respectively were chosen. The goal is to compare their performance and which features they deem as most important and finally statistically measure their performance.

### A. Random Forest

In Random Forest (RF) classification, the class that has earned the most votes from the individual trees in the ensemble is voted as the output class. This model incorporates two different mechanisms: Bagging and random selection of features [13]. The training dataset is divided into m subset randomly with replacement (bagging). Each subset of data is trained, and m total decision trees are created. To determine the optimal number of decision trees, features are selected in the random manner. Thus, every decision tree make is own decision and finally the individual with highest votes is considered as the output class [13]. RF classifier gives the one of the best performances among all existing ML tools. It can handle big dataset with great speed as it is parallelizable. Moreover, by applying this tool, the dataset with many features; 32 for "Road traffic Accident" dataset can be handled without any deletion of features. Since the 'Road traffic Accident' dataset has many missing values, RF is an appropriate model to handle the missing dataset as it still provides high performance with the missing dataset. And finally, for the 'Road traffic Accident' which has a highly unbalanced dataset, the methods like balancing the class weight for the target variables classes helps to resolve that problem [14]. One potential problem with RF model is it gives higher importance to attributes which have higher number of levels [13].

### B. XGBoost

According to Chang et al. [15], XGBoost uses the loss function's second-order Taylor expansion and a regularisation term to balance the model's complexity and the loss function's reduction. Thus, XGboost is highly successful in preventing overfitting and providing the optimal solution for the dataset. This model also provides parallelization thus reducing the executing time and increases the model precision [15].

### C. Explanatory data Analysis (EDA) and preprocessing

A synthetic dataset was created with make classification function form sklearn.datasets [14]. The synthetic dataset had total of 10000 rows and 4 different features. The target feature had imbalanced dataset with 8500 of 0 class and 1500 of 1 class.

The goal for creating the synthetic data was to test the hypotheis that the performance of RF is significantly different than the performance of XGBoost method. The synthetic data acts as a control in this experiment.

The 'Road Traffic Accidents' dataset was obtained from Kaggle [3]. The data was collected from Addis Ababa Sub city police departments and was published by Saurabh Shahane in Kaggle. The original dataset has 12316 instances and 32 different attributes. The target variable is Accident severity which have 3 classes: Slight injury, Serious injury and fatal injury. Our of 32 different attributes, there were missing values in ”Educational level”, ”Vehicle driver relation”, ”Driving experience”, ”Type of vehicle”, ”Owner of vehicle”, ”Service year of vehicle”, ”Defect of vehicle”, ”Area accident occurred”, ”Lanes or Medians”, ”Road alignment”, ”Types of Junction”, ”Road surface type”, ”Type of collision”, ”Vehicle movement”, ”Work of causality”, ”Fitness of causality” .There is no duplicated row found for this dataset.

The dataset was highly imbalanced with the target variable of slight injury with 10415 counts, serious injury with 1743 counts and fatal injury with 158 counts. In the dataset there were only two continuous variables; Numbers of Vehicle involved and Number of casualties. When the number of vehicles involved and the accident severity count was plotted, it showed that when the number of vehicles was 2, the largest number of casualties for slight, serious and fatal injuries were noted.

Since most of the variables with missing values were categorical, mode was used as the method of data imputation. From the heatmap for correlation matrix, the casualty class, sex of casualty, age band of casualty and casualty severity were found to be highly correlated. Since Random Forest and XGBoost have their own feature selection properties, I decided not to remove the correlated variables.The time variable was dropped as the format of the time didn't fit the analysis.

The dataset was splitted into test, train and validation set. The train and test dataset was split in 80:20 ratio. LabelEncoder was used to convert categorical values to integers. All the train validation and test dataset was scaled using Standard scaler. The whole training dataset (without feature selection) was fitted into Random Forest and XGBoost model. Recursive

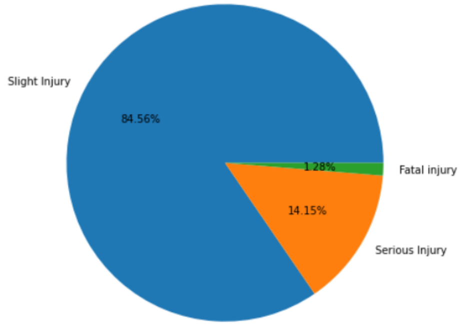


Fig. 1. Percentage of categories in Accident Severity

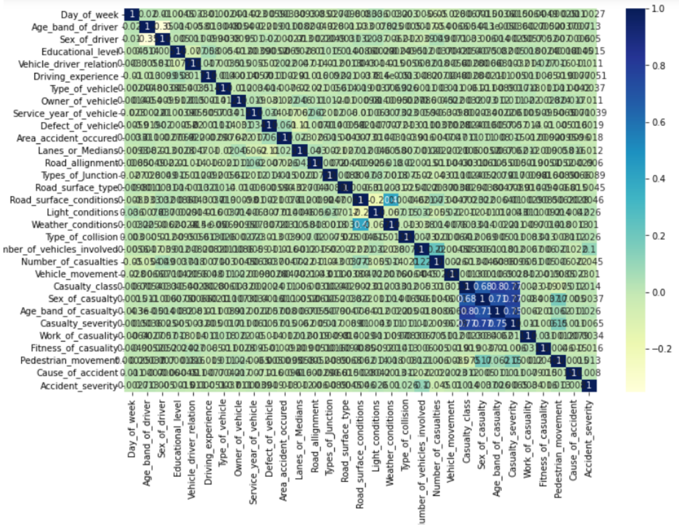


Fig. 2. Correlation matrix for different features

Feature Elimination (RFE) to find the most important variables. In RFE, present collection of features is trimmed of its least significant features. Once the appropriate number of features to pick has been attained, the technique is recursively repeated on the trimmed set [16].

The performance metric used for this analysis was Precision , recall and f1-score. Accuracy was not selected as the performance metric as the target variables has unbalanced ratio.

The randomized search CV was used to find the best parameters for both Random forest and XGBoost method. The class weight balance was used because it provided the class weight for the target variables according to the ratio of samples.

## IV. Experiments

Before analyzing the performance of Random Forest and XGBoost methods for the ‘Road Traffic Accidents' dataset, it was trained and tested in an imbalanced synthetic dataset. This was done because it provides some idea on how to models performs in other imbalanced dataset apart from the ‘Road Traffic Accident' dataset. As observed in table 1, the weighted average for F1 score was 0.93 and 0.92 respectively.

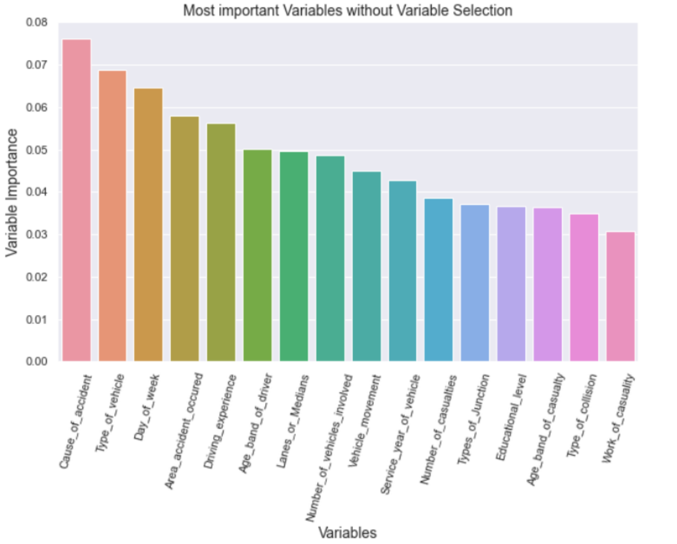


Fig. 3. Most important features obtained from RF using Recursive Feature Elimination

To statistically determine which model performed better, a 5x2cv paired t test was performed. From the analysis, the p-value obtained was 0.91 which was greater than the critical value of 0.05. Therefore, for the synthetic data, it could be concluded that both models have similar performance, and one cannot be said to have outperformed the other one.

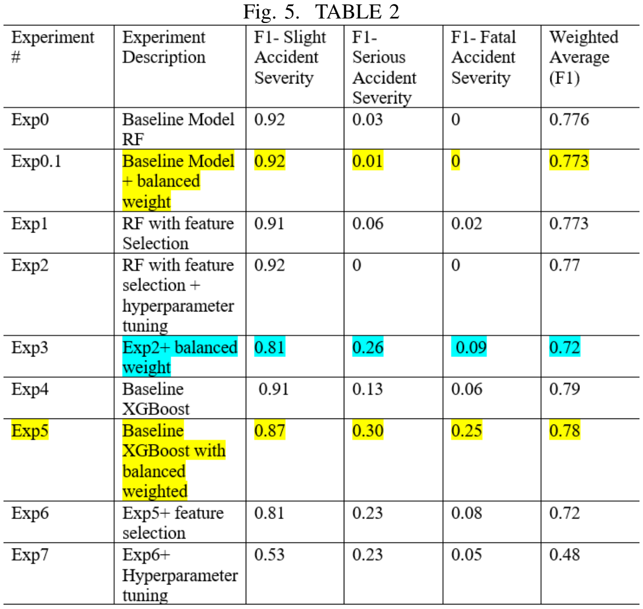
Fig. 4. TABLE 1

| Experiment # | Experiment Description | Class :0 | Class:1 | Weighted average |
|---|---|---|---|---|
| SYN RF | RF model with Synthetic data+ balanced weight | 0.96 | 0.75 | 0.93 |
| SYN XGBoost | XGBoost model with Synthetic data+ balanced weight | 0.95 | 0.73 | 0.92 |

Based on the results from experiments with synthetic data, a total of 9 experiments were performed for Random Forest model and XGBoost combined for the ‘Road Traffic Accidents' dataset. The baseline model for RF and XGBoost model were the one where the training data was trained and testing data was tested without any feature selection, hyperparameter tuning or balancing the weights of level in target features. The other experiments were the model with feature selection, model with balancing weights of levels for target feature, model with feature selection, hyperparameter tuning and balancing weights of levels for target features (table 2). Based on the results, it was found that all the RF experiments have similar weighted average F1 value of 0.77 except the experiment 3 where feature selection, hyperparameter tuning and balancing weight was applied. In Experiment 3, weighted average F1 value was 0.72 but the F1 value for less frequent class: serious and fatal accident severity was lot higher compared to other experiments. This is important because our model is able perform well for each class and not just the classes with highest number of instances i.e., Slight accident Severity. For the XGBoost model, Baseline model added with balancing the weights of level in target features has the best performance. This is because it has the weighted average of 0.78 which is among the highest. Moreover, it had higher number of f1 values for less frequent class i.e., serious, and fatal accident level for target feature.

When the statistical difference in performance (weighted F1value) was tested between baseline model with balancing weights of levels for target features (i.e., Exp0.1 and Exp5), using 2-fold 5 repeats paired t-test., the p-value was obtained to be 0.09 which was greater than the critical value of 0.05. Therefore, the performance of one model was not significantly better than other. On the other hand, when the model statistical difference in model performance (weighted f1) tested between Exp3 and Exp5 (models where f1 values for less frequent classes are higher), the p-value was obtained to be 0.015 which is smaller than the critical value of 0.05. This signifies that the XGBoost performs well when Experiment 3 and Experiment 5 are compared.

Fig. 5. TABLE 2



| Experiment # | Experiment Description | F1- Slight Accident Severity | F1- Serious Accident Severity | F1- Fatal Accident Severity | Weighted Average (F1) |
|---|---|---|---|---|---|
| Exp0 | Baseline Model RF | 0.92 | 0.03 | 0 | 0.776 |
| Exp0.1 | Baseline Model + balanced weight | 0.92 | 0.01 | 0 | 0.773 |
| Exp1 | RF with feature Selection | 0.91 | 0.06 | 0.02 | 0.773 |
| Exp2 | RF with feature selection + hyperparameter tuning | 0.92 | 0 | 0 | 0.77 |
| Exp3 | Exp2+ balanced weight | 0.81 | 0.26 | 0.09 | 0.72 |
| Exp4 | Baseline XGBoost | 0.91 | 0.13 | 0.06 | 0.79 |
| Exp5 | Baseline XGBoost with balanced weighted | 0.87 | 0.30 | 0.25 | 0.78 |
| Exp6 | Exp5+ feature selection | 0.81 | 0.23 | 0.08 | 0.72 |
| Exp7 | Exp6+ Hyperparameter tuning | 0.53 | 0.23 | 0.05 | 0.48 |

## V. Conclusion and Future Work

Based of experiments performed with synthetic data and 'Road Traffic Accidents' data, the model performance with both Random Forest and XGBoost model with default parameters were not different. However, XGBoost performed with default parameters and with balancing weights for target variables where the performance of less frequent target category: serious and fatal accident severity was significantly higher. In many empirical experiments, Random Forest have better performance compared to XGBoost which is contradictory to the results [17]. According to Lev [18], the reason for XGBoost to perform better for imbalanced dataset is if a tree fails to predict a class correctly, the subsequent calculations will provide more weight to the incorrect class. Usually, the incorrect class is the less frequent target label class. Thus, this build in mechanism of boosting in XGBoost is better suited to handle the imbalanced dataset.

Further studies are required on which classification method is better for multiclass imbalanced dataset. Thus, many experiments need to be performed with different hyperparameter tunings and diverse dataset. Moreover, mathematical basis for better performance of RF or XGBoost for imbalanced dataset need to be studied.

In the future, I wish to apply neural networks and identify if the performance of such ML tool is better for 'Road Traffic Accidents' dataset when compared to ensemble methods. Moreover, I plan to explore the interpretability of such neural networks.

## References

[1] "Road traffic injuries," World Health Organization, 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. [Accessed: 23-Dec-2022].

[2] A. Ditcharoen, B. Chhour, T. Traikunwaranon, N. Aphivongpanya, K. Maneerat and V. Ammarapala, "Road traffic accidents severity factors: A review paper," 2018 5th International Conference on Business and Industrial Research (ICBIR), 2018, pp. 339-343, doi: 10.1109/ICBIR.2018.8391218.

[3] T. T. Bedane, "Road traffic accident dataset of Addis Ababa City," Mendeley Data, 02-Nov-2020. [Online]. Available: https://data.mendeley.com/datasets/xytv86278f/1. [Accessed: 23-Dec-2022].

[4] X. Jianfeng, G. Hongyu, T. Jian, L. Liu, and L. Haizhu, "A classification and recognition model for the severity of road traffic accident," Advances in Mechanical Engineering, vol. 11, no. 5, p. 168781401985189, 2019.

[5] L. Auria and R. A. Moro, "Support Vector Machines (SVM) as a technique for solvency analysis," SSRN Electronic Journal, 2008.

[6] A. Raj, "Everything about support vector classification - above and beyond," Medium, 30-Mar-2022. [Online]. Available: https://towardsdatascience.com/everything-about-svm-classification-above-and-beyond-cc665bfd993e. [Accessed: 23-Dec-2022].

[7] P. Ranganathan and R. Aggarwal, "Common pitfalls in statistical analysis: Linear Regression Analysis," Perspectives in Clinical Research, vol. 8, no. 2, p. 100, 2017

[8] L. G. Cuenca, E. Puertas, N. Aliane and J. F. Andres, "Traffic Accidents Classification and Injury Severity Prediction," 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE), 2018, pp. 52-57, doi: 10.1109/ICITE.2018.8492545.

[9] H. Chen, S. Hu, R. Hua, and X. Zhao, "Improved naive Bayes classification algorithm for Traffic Risk Management," EURASIP Journal on Advances in Signal Processing, vol. 2021, no. 1, 2021.

[10] "What is neural network?," AWS, 2022. [Online]. Available: https://aws.amazon.com/what-is/neural-network/. [Accessed: 22-Dec-2022].

[11] D. Lei, X. Chen , and J. Zhao, "Opening the black box of deep learning," arXiv, May 2018.

[12] T. G. Dietterich, "Ensemble Methods in Machine Learning," Multiple Classifier Systems, pp. 1–15, 2000.

[13] M. Ferreira (2022). Ensamble Learning [PDF slides]. https://dal.brightspace.com/d2l/le/content/230475/Home

[14] F. Pedregosa, E. Duchesnay, M. Perrot, M. Brucher, D. Cournapeau , A. Passos, J. Vanderplas, V. Dubourg, R. Weiss , P. Prettenhofer, M. Blondel , O. Grisel, B. Thirion, V. Michel , A. Gramfort, and G. Varoquaux, "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.

[15] W. Chang, Y. Liu, Y. Xiao, X. Yuan, X. Xu, S. Zhang, and S. Zhou, "A machine-learning-based prediction method for hypertension outcomes based on medical data," Diagnostics, vol. 9, no. 4, p. 178, 2019.

[16] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," Machine Learning, vol. 46, no. 1/3, pp. 389–422, 2002.

[17] Y. Li and W. Chen, “A comparative performance assessment of ensemble learning for credit scoring,” MDPI, 13-Oct-2020. [Online]. Available: https://www.mdpi.com/2227-7390/8/10/1756. [Accessed: 26-Dec-2022].

[18] A. Lev, “XGBoost versus Random Forest,” Qwak’s Blog, 19-Dec-2022. [Online]. Available: https://www.qwak.com/post/xgboost-versus-random-forest. [Accessed: 26-Dec-2022].