# Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach

SCH Univ.
Dept. of AI and Bigdata
Sunghun Kim
Developing Traffic Safety

# contents


1. Introduction
2. Previous literature review
3. Methodology
4. Experiment, evaluation, and discussion
5. Experimental result analysis

# 1. Introduction

# 1. Introduction

1. RTA (road traffic accident)
    - Churning the world with killing thousands and bringing demolition of property in a day without discrimination
    - Does not give much attention to mitigate the severity
    - Not occur by chance, it has patterns and can be predicted and avoided

2. Getting insights and identify the underlying cause of vehicle accidents and related factors

↓

**Reduce road traffic accidents**

# 2. Previous literature review

# 2. Previous literature review

Conventional statistical-based approach lacks the capability to deal with multidimensional datasets

To address the limitations of traditional models, many studies used ML approach due to its predictive supremacy, time consuming

<State-of-the-art model for accidents>

- K-means
- SVM (Support vector machine
- KNN (K-Nearest Neighbors)
- DT (Decision Tree)
- ANN (Artificial Neural Network)
- CNN (Convolution Neural Network)
- LR (Logistic Regression)

## Road Accident Analysis by Kwon et al.

- Model Used: Naïve Bayes and Decision Tree
- Methodology: Binary Regression for Performance Comparison
- Finding: NB showed higher sensitivity to risk factors compared to DT

---

## Road Accident Analysis by Sharma et al.

- Model Used: Support Vector Machine(SVM) and Multi Layer Perceptron(MLP)
- Independent Variables: Alcohol and Speed considered as key factors
- Methodology: Model Comparing by Accuracy
- Finding: SVM with RBF kernel achieved higher accuracy (94%) compared to MLP (64%)

## Motorcycle Crash Analysis by Wahab and Jiang

- Data: crash accidents in Ghana
- Model used: MLP, PART and SimpleCART
- Methodology: Used Weka tools to compare the model and applied InfoGainAttributeEval to see the most influential variable for motorcycle crash
- Finding: SimpleCART model showed better accuracy than other classification models

# 3. Methodology

K-means clustering + Random Forest

For creating new features

Classifier

## 3.1 Road accident dataset manipulation

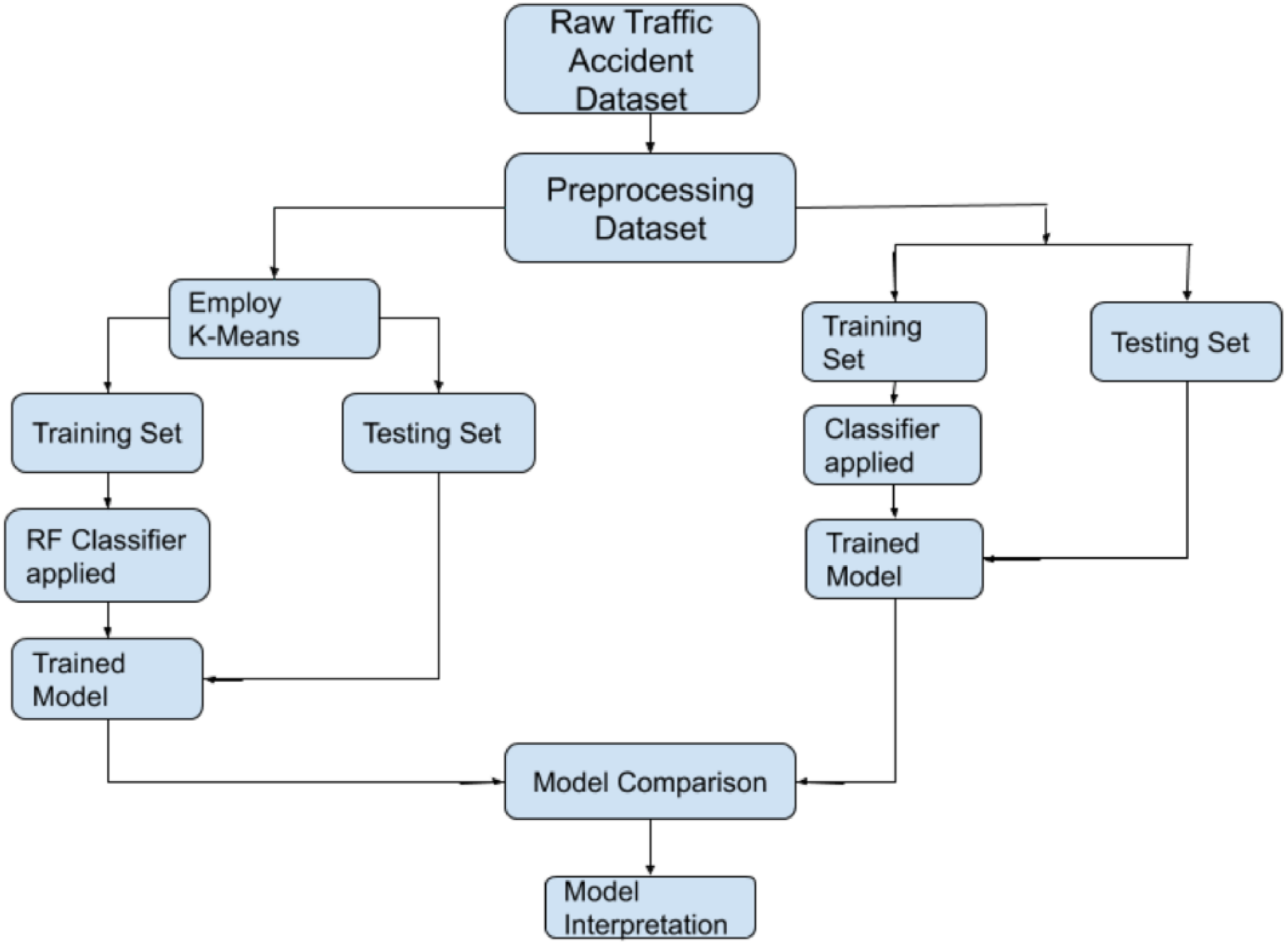



**Fig. 1** Flowchart of proposed model framework for predicting road traffic accident—case of Ethiopia

# 3.1 Road accident dataset manipulation (Data)

Raw traffic accident dataset

- 5000 road traffic accidents collected from federal traffic police agency
- 2011 to 2018 in Addis Ababa

<Data Feature>

- Accident time
- Driver age
- Sex
- Driver experience
- Type of vehicle
- Service year
- Location
- Road condition
- Light condition
- Weather condition
- Casuality class
- Casuality age
- Casuality sex
- Severity

# 3.1 Road accident dataset manipulation (Preprocessing / Data Splitting)

Data preprocessing

- Data cleaning
- Missing value handling
- Outlier treatment
- Dealing with absolute value → encoding and normalization

→ Prediction model

Data Splitting

- 70% train data, 30% test data

## 3.2 K-means techniques

Unobserved Heterogeneity

- Unobserved characteristics associated with observed characteristics during model building

K-means

- Effective clustering maintains similarity within clusters and diversity between them
- Create new features
- Combined with classification, enables swift, accurate training, and reduced computational memory usage

## 3.2 K-means techniques

K-means algorithm

1. Randomly initialize and select the Cj-centroids
2. Calulate the distance between each instance to the Cj-centroid
3. Compute mean of each data points in each cluster to find their centroid
4. Repeat the forementioned steps until each points assigned to their nearest cluster

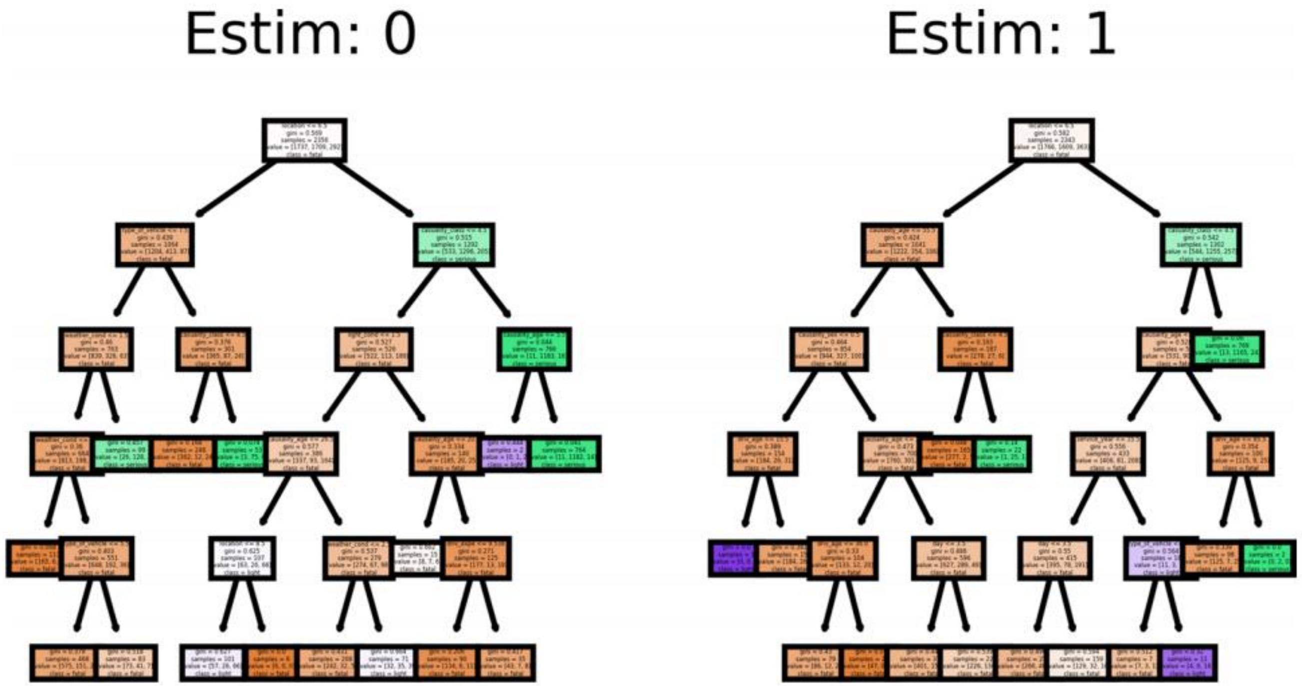
**Squared error function**

$$f(x) = \sum_{i=1}^{k}\sum_{j=1}^{n}\left|X_i - C_j\right|^2$$

# 3.3 Random forest techniques

- Decision trees prone to overfitting → Random Forest mitigates using multiple trees
- Robust algorithm for large datasets (provides accurate predictions)
- Maintains accuracy with missing data

Fig. 2 Sample random forest (n-estimator= 5)




SN Applied Sciences
A SPRINGER NATURE journal

# 4. Experiment, evaluation, and discussion

# 4.1 Data manipulation

## Missing value handling

- Ignore or drop missing value
- Fill using different method

  ↪ numeric variables: mean / categorical variables: mode

## Categorical Value Encoding

- Machine learning require numeric values to predict a model
- Among 14 variables, 10 of them are categorical values
- Predictive and target variables converted into numeric using one-hot-encoding and label encoding

| | Missing Values | % of Total Values |
|---|---|---|
| service_year | 1128 | 22.6 |
| driv_expe | 898 | 18.0 |
| type_of_vehcle | 571 | 11.5 |
| driv_age | 538 | 10.8 |
| sex | 422 | 8.5 |
| causality_age | 320 | 6.4 |
| location | 315 | 6.3 |
| causality_sex | 166 | 3.3 |
| light_cond | 157 | 3.2 |
| day | 131 | 2.6 |
| casuality_class | 110 | 2.2 |
| road_cond | 105 | 2.1 |
| severity | 74 | 1.5 |
| weather_cond | 1 | 0.0 |

# 4.2 Evaluation metrics

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$f1 - score = 2 \times \frac{(Precision \ \times Recall)}{(Precision + Recall)}$$

- TP: it shows predictive is positive and it is normally true
- TN: it implies predictive is Negative and it is normally True
- FP: denotes predictive is positive and it is normally false
- FN: represents predictive is negative and it is false

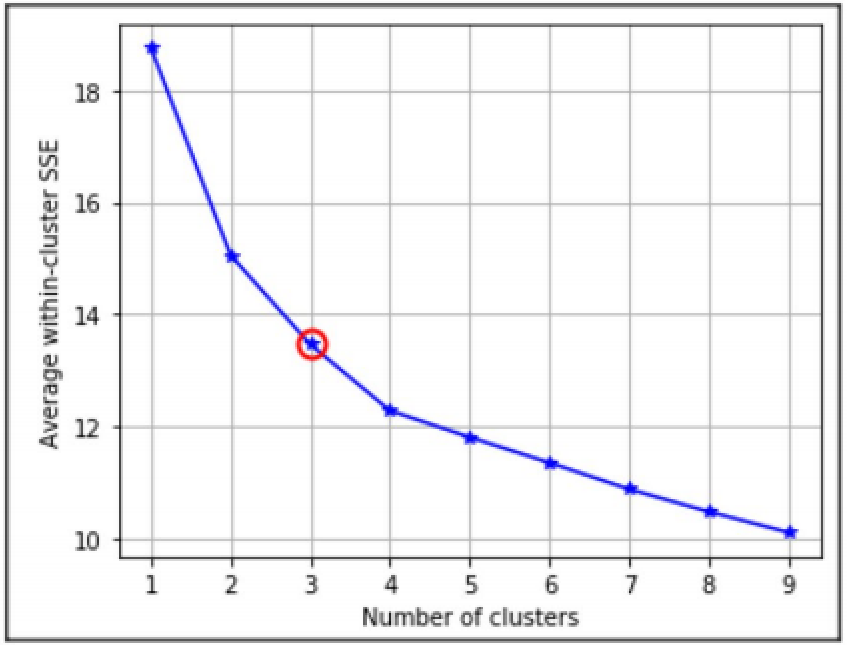
# 5. Experimental result analysis and discussion

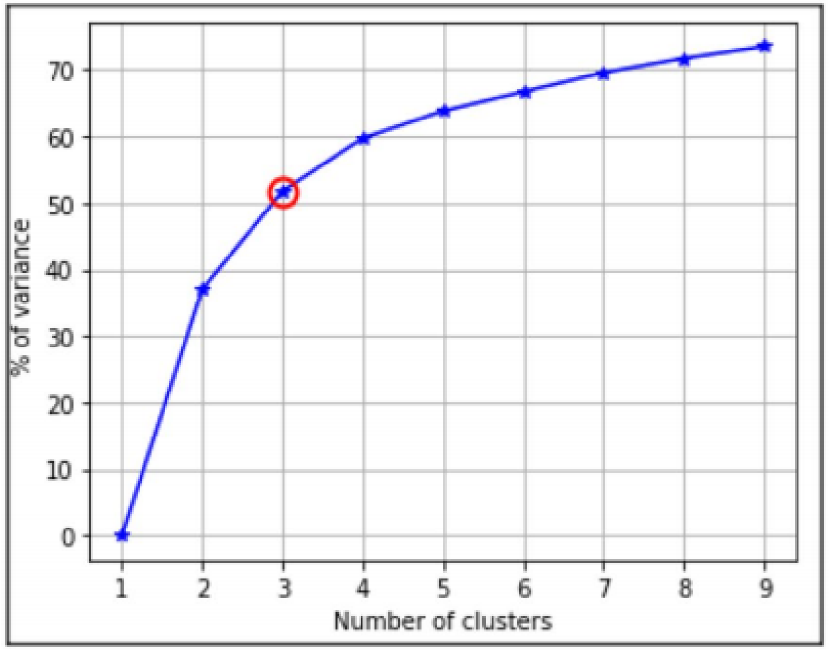
# 5.1 Choosing K

- No specific solution to find the exact value of K
- K increases, the sum of squared distance leans towards zero and the percentage of variances increase

Inertia

- Sum of squared distances
- Sum of distances between data points and cluster centroids

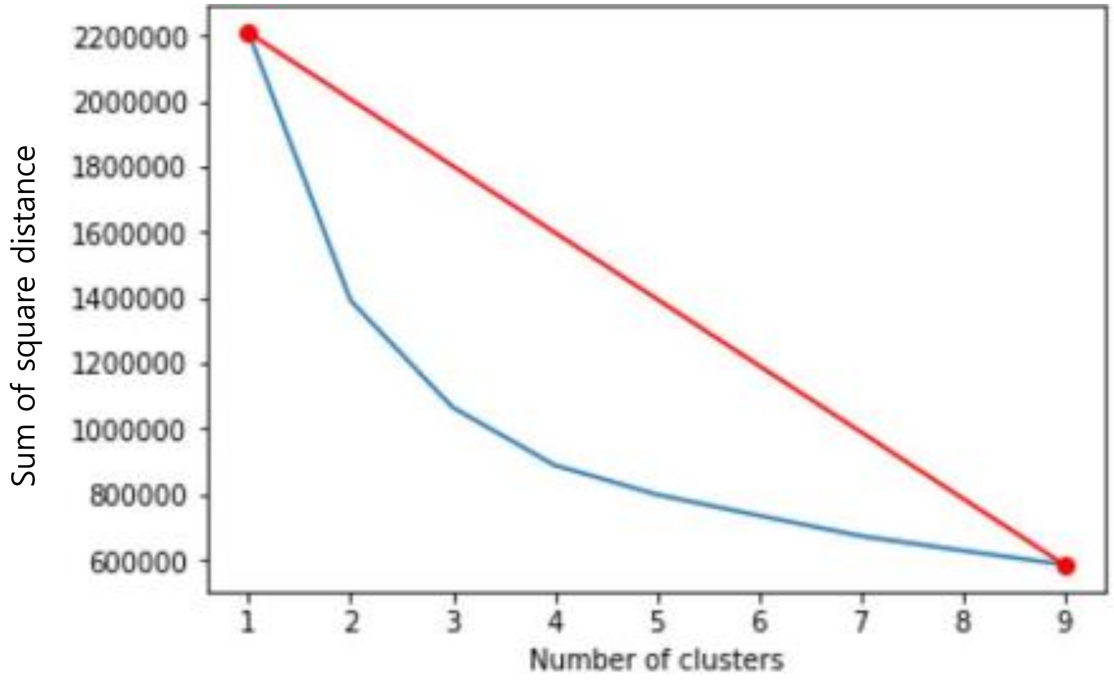


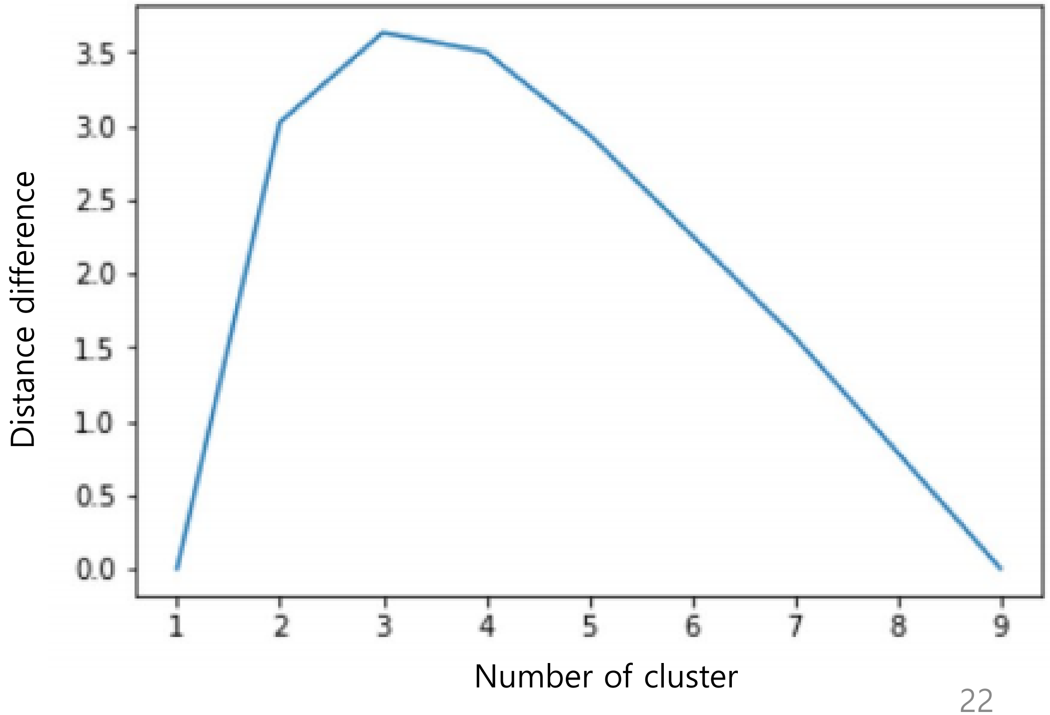



## 5.1 Choosing K

1. Based on elbow method, the elbow resembles a suitable 'k' value.
2. Due to ambiguity, a line connecting 'k' values 1 and 9 was drawn.
3. Optimal 'k' was deduced from the point where this line maximized distance from the original function.
4. Consequently, 'k' was determined to be 3 for effective clustering analysis.

→ Road accident dataset clustered into three groups



## 5.3 Model performance evaluation

**Table 1** Performance evaluation of classifiers and proposed approach

| S. No | Classifier | Testing set without new feature | | | | Testing set with new feature | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | f1 score | Accuracy | Precision | Recall | f1 score | Accuracy |
| 1 | K Means | 47 | 42 | 43 | 42.25 | 36 | 36 | 35 | 35.83 |
| 2 | LR | 85 | 87 | 84 | 86.83 | 99 | 99 | 99 | 99.13 |
| 3 | RF | 86 | 88 | 87 | 87.77 | 100 | 100 | 100 | 99.86 |
| 4 | SVM | 69 | 68 | 65 | 68.45 | 76 | 73 | 70 | 73.13 |
| 5 | KNN | 64 | 65 | 62 | 64.97 | 68 | 69 | 66 | 68.58 |

**Table 2** The execution time of models (ms)

| Model | Training time | Testing time |
|---|---|---|
| K-means | 191 | 2.57 |
| LR | 231 | 1.29 |
| RF | 399 | 38 |
| SVM | 566 | 134 |
| KNN | 9.7 | 87 |
| K-means-RF | 295 | 5.71 |

## 5.3 Model performance evaluation

**Table 3** Performance comparison of related work models

| References | Classifier | Dataset | Accuracy |
|---|---|---|---|
| Gu et al. [21] | PSO-SVM | China | – |
| Xiao et al. [52] | SVM, KNN (Ensemble) | I-880 data set | 99.33% |
| Castro et al. [15] | BN, JR8 and MLP | DVSA—UK | 72.39%, 72.02%, 71.70% Respectively |
| Al-Radaideh et al. [4] | RF, ANN (backpropagation), SVM | Uk | 80.6%, 61.4%, 54.8% respectively |
| Casado et al. [14] | LCC, MNL | Spain | – |
| Wahab et al. [51] | MLP. SimpleCart, PART | Ghana | 72.16%, 73.45%, 73.81% respectively |
| Sameen et al. [40] | MLP, BLR, RNN | Malaysia | 65.48%, 58.30%, 71.77% respectively |
| Fentahun [18] | J48, ID3, PART | Ethiopia | 81.21%, 81.01%, 81.18% |
| Seid et al. [42] | HMR | Ethiopia | NA |
| Abebe et al. [1] | DSA | Ethiopia | – |
| Lytin et al. [30] | UBA | Ethiopia | – |

# 5.4 ANN experiment analysis

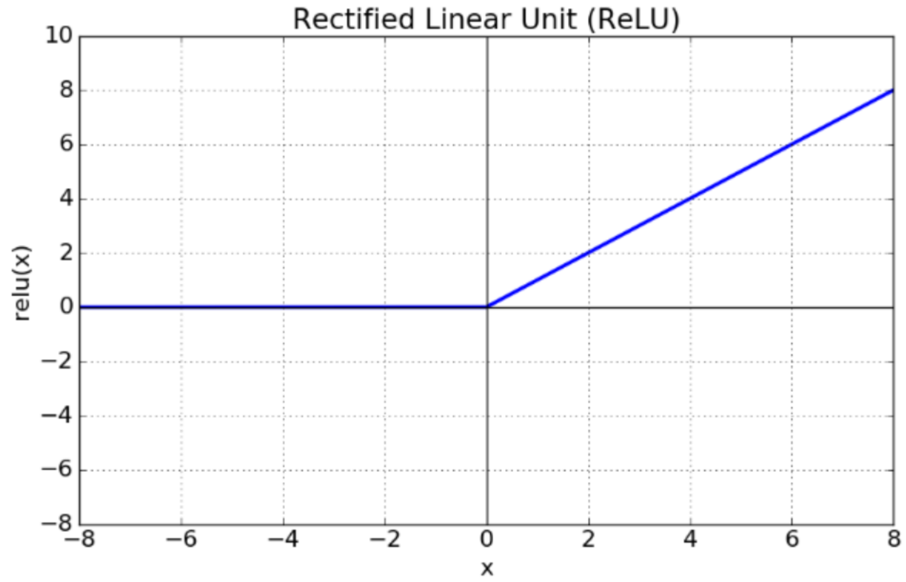
- Input layer → Rectifier activation function
- Output layer → Sigmoid activation function

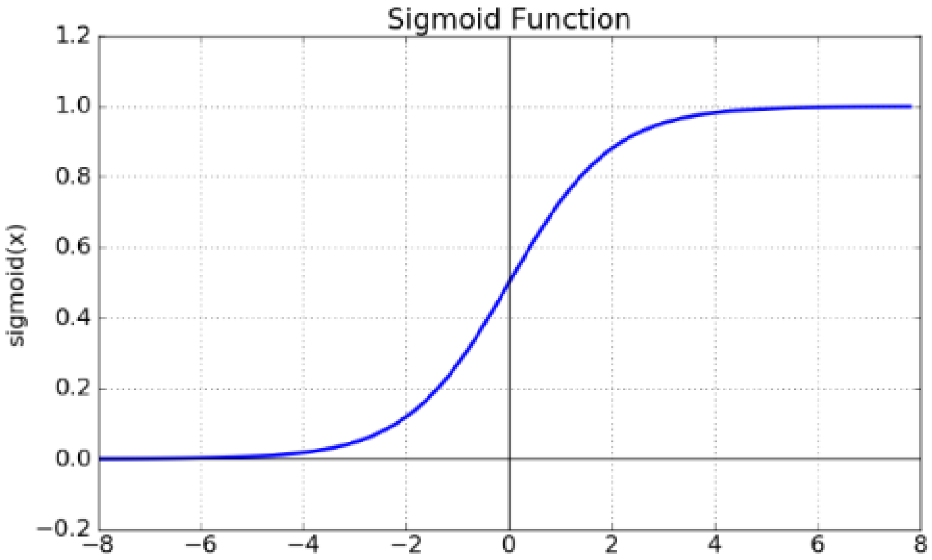
**Table 4** Test accuracy, loss, and ROC curve value of ANN model with multiple dense layers

| Model | Dense layer | Test accuracy (%) | Test loss | ROC curve (%) |
|---|---|---|---|---|
| $Model_1$ | 2 | 88.77 | 0.3819 | 96.1 |
| $Model_2$ | 3 | 88.77 | 0.3622 | 96.1 |
| $Model_3$ | 4 | 88.03 | 0.3686 | 96.1 |

**Table 5** Comparison of ANN and proposed model performance with different metrics (%)

| Model type | Precision | Recall | F1 score | Accuracy |
|---|---|---|---|---|
| ANN | 88 | 88 | 88 | 88 |
| Proposed model | 100 | 100 | 100 | 99.86 |




https://m.blog.naver.com/PostView.naver?isHttpsRedirect=true&blogId=handuelly&logNo=221824080339

# 5.4 ANN experiment analysis

**Fig. 7** The validation and loss accuracy of different ANN Model

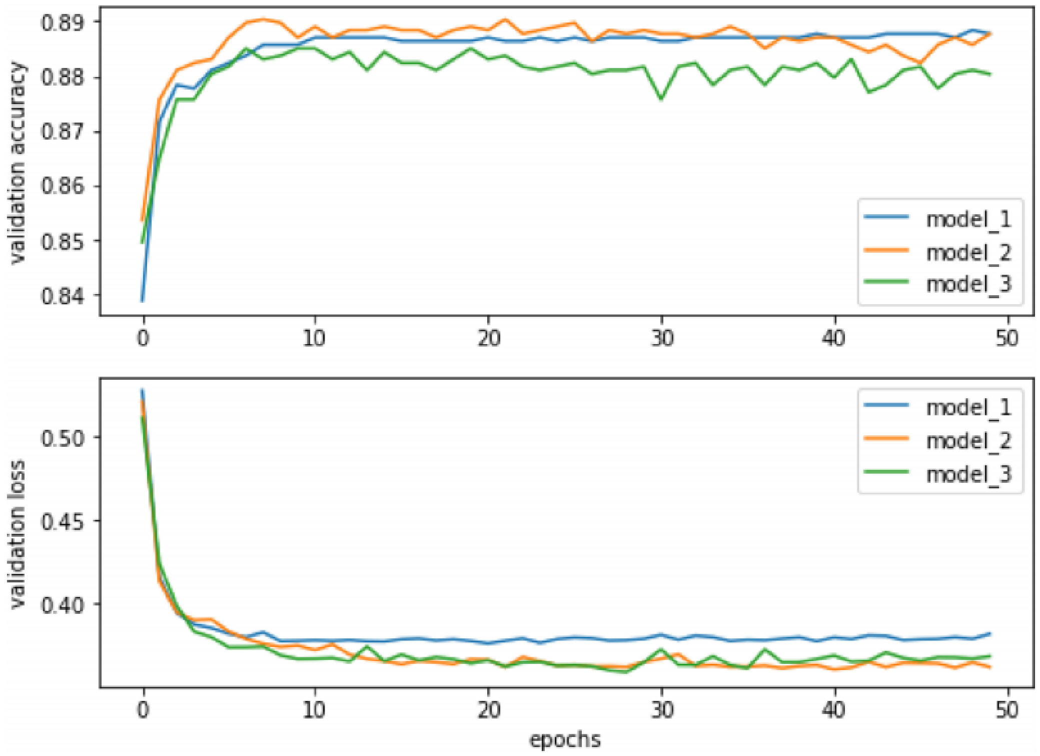
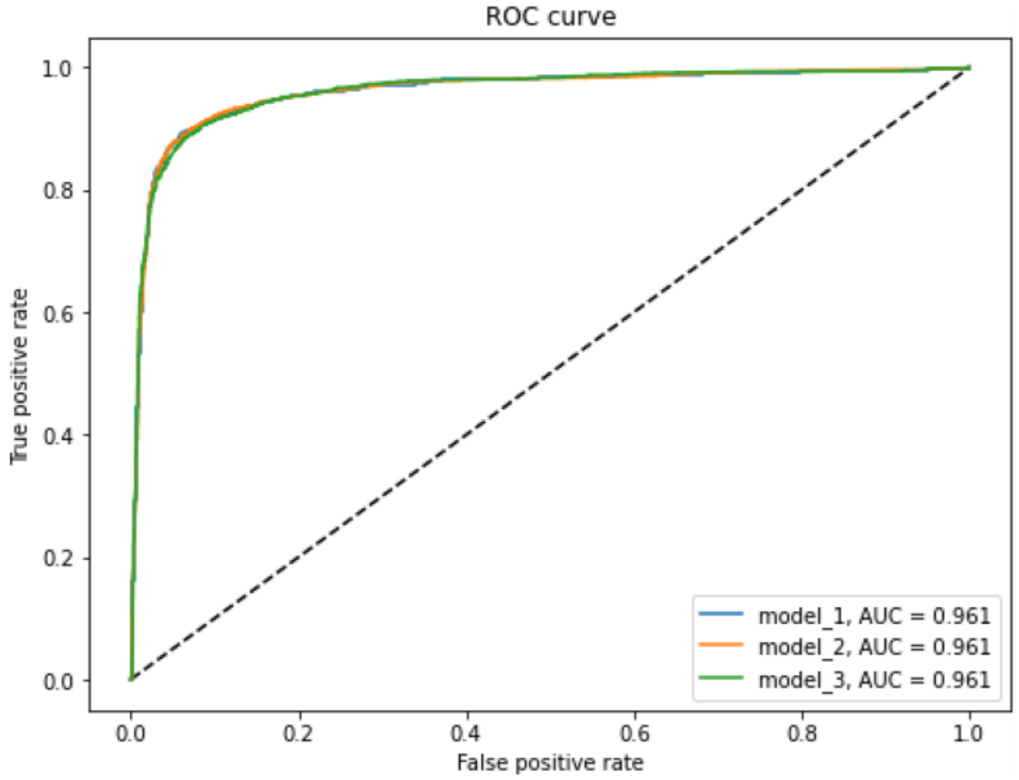



**Fig. 8** ROC curve of different ANN models



# 5.6 Random forest interpretation

1. Decision Tree

$$f(x) = Cfull + \sum_{k=1}^{M} contrib(x,k)$$

$Cfull$ : Root node value

$M$ : Number of leaves in the tree

$contrib(x,k)$ : kth feature contribution in feature vector $x$

---

2. Random forest predict function

$$g(x) = \frac{1}{J}\sum_{j=1}^{J} f_j(x)$$

$$g(x) = \frac{1}{J}\sum_{j=1}^{J} C_j full + \sum_{k=1}^{M}(\frac{1}{J}\sum_{j=1}^{J} contrib_j(x,k))$$

$J$ : Number of decision tree

$f_j(x)$ : Prediction functions for each tree

# 5.6 Random forest interpretation

Serious injuries

- day
- Location
- Causality age
- Driver experience
- Light condition
- Casualty sex
- Type of vehicle

---

Minor injuries

- Light condition
- Causality class
- Causality age
- Causality sex

---

Fatal accident severity

- Driver age
- Service year
- Weather condition
- Casualty class

# 6. Conclusion

**Hybrid Approach Superiority**

- Developed method outperforms traditional machine learning methods for RTA dataset severity prediction.

**K-Means Integration**

- Utilized K-Means clustering integrated with Random Forest classification, showing superior performance over other models. (99.86% accuracy)

**Target-Specific Insights**

- Highlighted the effectiveness of combining Clustering and Classification to identify key factors for different accident severity classes.

**Future Prospects**

- Aiming to strengthen model efficacy by exploring additional datasets for further insights and improved accuracy.

# Thank You