Context Aware Ontology based Hybrid Intelligent Framework for Vehicle Driver Categorization

Sohail Sarwar, Saad Zia, Zia ul Qayyum, Muddesar Iqbal, Muhammad Safyan, Shahid Mumtaz, Raul Garcia, Kostromitin, K. I.

Castro ***Abstract*— In public vehicles, one of the major concerns is driver's level of expertise for its direct proportionality to safety of passengers. So before a driver is subjected to certain type of vehicle, he should be thoroughly evaluated and categorized with respect to certain parameters instead of only one-time metric of having driving license. These aspects may be driver's expertise, vigilance, aptitude, experience years, cognition, driving style, formal education, terrain, region, minor violations, major accidents and age group etc. The purpose of this categorization is to ascertain suitability of a driver for certain vehicle type(s) to ensure passengers' safety. Currently, no driver categorization technique fully comprehends the implicit as well as explicit characteristics of drivers dynamically. In this paper, machine learning based dynamic and adaptive technique named *D-CHAIT* (Driver Categorization based on Hybrid Artificial Intelligence Techniques) is proposed for driver categorization with an objective focus on driver's attributes modeled in DriverOntology. A supervised mode of learning has been employed on a labeled dataset, having diverse profiles of drivers with attributes pertinent to drivers' perspectives of demographics, behaviors, expertise and inclinations. A comparative analysis of *D-CHAIT* with three other machine learning techniques (Fuzzy Logic, Case Based Reasoning, and Artificial Neural Networks) is also presented. The efficacy of all techniques was empirically measured while categorizing the drivers based on their profiles through metrics of accuracy, precision, recall, f-measure performance and associated costs. These empirical quantifications assert *D-CHAIT* as a better technique than contemporary ones. The novelty of proposed technique is signified through preprocessing of feature attributes, quality of data, training of machine learning model on more relevant data and adaptivity.**

***Index Terms*— Artificial Neural Networks, Case Based Reasoning, Vehicle Driver Categorization, Fuzzy Logic, Machine Learning**

I. Introduction

The foremost concern of any transportation authority would be assuring the safety of passengers using their facility. Adoption of modern technologies in transportation has not only enhanced the degree of comfort for travelers but causing serious threats to passengers. Such threats are evident from 43,000 causalities just in year 2005 [1]. Out of 1.2 million accidents, 20-30% were caused by negligence of drivers. This gives rise to the need of thoroughly assessing driver's expertise in certain terrains, abilities to drive on longer routes with certain vehicles, violation history, and regions etc. Once the picture of the driver's ability is clear by rightly categorizing the drivers, he can be delegated to certain vehicle, region and route.

Different techniques have been proposed for driver categorization [1, 2, 3, 4, 5] that categorize driver profiles by exploiting the "unsupervised mode" of learning for classification. Some techniques take into account only driver's behavior for categorizing the drivers as given in [1, 2]. Others consider only the environmental factors which distract the drivers (mobile phones, conversing with passengers, turning patterns and lane changing) using Hidden Markov Models (HMM) for better future measures [4]. Some techniques evaluate the drivers over a specific temporal scale that does not fully depict the category of driver in normal circumstances. Also few techniques consider certain age groups [5] for their studies. The driving patterns of vehicle are used to classify drivers based on their reactive-ness to uncertain situations [6]. Similarly, *adaptive cruise control* and *lane keeping* have been incorporated in autonomous vehicles [7] by simulating behavior of drivers without covering all scenarios pertinent to driver's profiles. These techniques may not fully comprehend the characteristics (driving style, aptitude and personal details etc) of driver in categorization; specifically. Few driver categorization techniques such as based on CNN [7] are cost intensive, require a lot of training data and after categorizing the drivers, do not take advantage of reusing information of newly categorized drivers for future classifications. This prevents machine learning techniques from being dynamic and adaptive to cater new scenarios dynamically. Lastly, few techniques [9, 10] claim to target the semantic web but formal and explicit descriptions of drivers using ontologies seem missing.

Keeping the facts above in view, an adaptive and dynamic

This work is submitted on 30 Nov 2018. We hereby acknowledge the funds support under the project PrivSoft sponsored by Higher Education Commission under the grant number:112116-Eg043.

Dr Sohail Sarwar is with the Department of Computing University of Gujrat Pakistan (e-mail: sohail.sarwar@seecs.edu.pk).

Saad Zia is with National University of Computing and Emerging Sciences (NUCES)-FAST, Pakistan
(emai: Saad.zia@nu.edu.pk)

Dr Zia Ul Qayyum is working with Allama Iqbal Open University Islamabad, UK (email: vc@uog.edu.pk)

Dr Muddessar is working with London South Bank University, UK (m.iqbal@lsbu.ac.uk, m.iqbal@essex.ac.uk) . He is also working as Visiting Senior Lecturer at University of Essex, UK.

Dr Muhammad Safyan is working with Government College University of Lahore Pakistan (e-mail: m.safyan@seecs.edu.pk)

Dr Raul Garcia Castro is with Universidad Politecnica de Madrid, Spain, (e-mail: rgarcia@fi.upm.es).

Dr. Shahid Mumtaz is working with Instituto de Telecomunicações
P-3810-193 AVEIRO – PORTUGAL (email: smumtaz@av.it.pt)

Kostromitin, K. I. is working with South Ural State University, 76, Lenin Avenue, Chelyabinsk, 454080, Russian Federation (email kostromitinki@susu.ru:)

driver categorization technique named *D-CHAIT* (*Driver Categorization through Hybrid of Artificial Intelligence Techniques*) is presented. The proposed technique, exploiting the notion of hybrid machine learning techniques for driver categorization, by modeling driver profiles through an ontology. It is dynamic enough to build a driver's profiles automatically with implicit parameters from real time data sources and explicit parameters acquired from the driver. The profiles of drivers are modeled by considering demographic, behavioral, and inclinatory aspects of the driver in an ontology named *DriverOntology* to benefit from semantic web technologies. After building the profile of drivers using Protégé in ontology, the proposed technique (*D-CHAIT)* classifies the drivers by exploiting the retrieval phase of Case Based Reasoning (CBR) and employs Artificial Neural Networks (ANN) in adaptation. Moreover, it updates the data repository containing driver information. This aids in dynamically reusing the profile of existing drivers in classifying upcoming drivers. Besides classifying the drivers through *DCHAIT*, another goal of our work is to recommend the most appropriate one among four Machine Learning (ML) techniques for driver categorization by making a comparative analysis in terms of performance and cost. A comparative analysis of *D-CHAIT* along with other ML techniques such as Case Base Reasoning (CBR) [9], Artificial Neural Networks (ANN) [10] and Fuzzy Logic (FL) [11] is also presented for categorizing the drivers. Drivers are categorized into one of the categories of *'Novice', 'Easy', 'Proficient' or 'Expert'* based on their profiles. Here it is worth mentioning that these driver categories were devised after a survey from the drivers, driving instructors, driver evaluation from the behavioral and cognitive perspectives. It is worth mentioning that some of the techniques used in academics for learner categorization have been consulted from literature [2, 3, 8, 9, 10, 13, and 15].

The rest of the paper is organized as follows: Section 2 provides an insight into the efforts accomplished for categorization of drivers. A view of ML techniques coupled with driver categorization is presented in section 3. Implementation technologies, content model and details are discussed in section 4. Section 5 presents the results and elaborates discussion from different perspectives of performance metrics while giving a direction of future research initiatives.

II. Literature Survey

A thorough literature survey has been carried out in order to have an idea of prevalent techniques for driver categorization, their advantages and the pitfalls to improve. These techniques have been spawned from different perspectives of machine learning such as data mining based techniques, Genetic Algorithms (GA), unsupervised classifiers, supervised classification predictors and others targeting web 3.0 for classification.

One of the major reasons for vehicular accidents in public transport is the inexperience drivers, external distractions disrupting the drivers attention, and suitability of drivers for certain vehicle types [4]. This paper focuses on categorizing the drivers based upon their behavior and environmental distractions through signals transmitted by CAN-Bus using Hidden Markov Models (HMM) and Gaussian Markov Models (GMM). It is claimed that an accuracy of almost 70% has been achieved in classifying the driver actions with 30% accuracy in identifying the distractions to be avoided. These initial experiments have revealed an encouraging degree of results for using CAN-Bus transmissions for classifying behavior of drivers and task distractions.

A thorough analysis and comparison of drivers in different age groups and experience years has been presented in [5]. The baseline hypothesis is to present lousiness of drivers in younger age group that is observed to increase after 3-4 experience years compared with careful behavior in first year of acquiring the license. The experiments are based upon relatively smaller sets of dataset samples appearing to be biased in some scenarios. It has been caused due to differing attitudes before and after such as knowledge of traffic rules/pitfalls, safety measures, and volunteer work. Another interesting finding is the emergence of risky driving patterns on weekends was observed compared with weekdays compared with first year and fourth year of driving. It is worth mentioning that young drivers adhere to safety guidelines when directed with specific feedback. Moreover, data representation and aggregation exploited for effective analysis of data were robust due to IVDR technology of behavioral pattern recognition of drivers.

In [8], driver's observable actions have been mapped over the anticipated actions (not observed in prior actions). For example, action of "changing lane" termed as the process of "mind-tracking". A number of cognition models were developed for assessing and categorizing the behavior of drivers.

Data mining, especially educational data mining [12, 13] termed as an emerging discipline, is claimed to have a great room for developing methods and exploring unique types of data that come from educational settings. Using these methods has potential to facilitate better understanding of contents for drivers.

A supervised mode of learning was employed in [6] to model and categorize the driving patterns. These patterns and behaviors are identified based upon certain parameters of vehicles around the "subject" vehicle. The parameters are termed as sets of states and actions. This approach may assist not only for safety of vehicles but also to maintain the degree of velocity and mobility. A variation of Support Vector Machine (SVM) was used for training and classifying the profiles of drivers in certain scenarios.

Lane-level localization for intelligently managing the lane changing in autonomous vehicles by capturing the imagery through GPS is discussed in [7]. Also, driver's behavior classification has been carried out through support vector machines keeping in view the lane changing patterns. The authors lay foundation of work based upon certain assumption such as prominently marked lanes, noisy interference of other vehicles to be discarded and availability of updated digital map. Almost all the aspects for training as well as testing of SVM classification were exploited in order to match the patterns to resolve future problems in autonomous vehicles.

Also, Convolutional Neural Network (CNN) was used as a baseline structure for SVM experiments.

Some other analogous techniques in learner classification can be used to categorize drivers are given in [14, 15, 16]. An unsupervised classification technique, linear regression, is used for modeling the quantity of accumulated knowledge pertinent to a learner. It uses variables linked to the learning activity, user experience and accumulated knowledge. Another analogy that can be used in driver categorization comes from domain of e-learning i.e. categorization of learners performed at concept level [17] that in turn is evaluated based on percent concepts covered in knowledge. After assessments, learners are classified based upon two aspects intellectually i.e. quality of answers and the time consumed in answering. Similarly a Supervised mode of learning has been used for sorting the slow learners out via performance prediction using a Naïve Bayes classifier [18]. This technique marks the students requiring remedial learning activities to be designed by an instructor. Results predict that Multi Layer Perceptron (MLP) [19] with 75% predictive accuracy has better performance than the rest of the techniques. The adaptivity of contents based upon the learner profile is discussed for recommending suitable contents using Random forest for classifiers.

Connected vehicular technologies have potential to assist in efficient management of traffic flow. However, changes in driver's behavior simulated by external conditions need to be observed for opting corrective measures. In [20], an automated system has been proposed for vehicle's safety based on driver characteristics i.e. psychological and demographic behavior of drivers along with speed/velocity and road condition with respect to surrounding vehicles.

The prevalent techniques used for driver' categorization discussed above exploit different mechanisms of the machine learning realm. However, each of these techniques may have associated issues with their usage in categorizing the drivers as discussed further. Data mining techniques are applied using unsupervised mode of learning and classification; such solution may not be handy when we want to be specific about categories while categorizing the drivers. Moreover, these techniques are not fragile enough to implicitly consider and cater new scenarios in the dataset for future decision making pertinent to the classification of the drivers. Techniques driven by a supervised mode of classification categorize the driver based upon behavioral aspects and may not fully comprehend the performance and inclinatory aspect as desired.

D-CHAIT targets to address the stipulated issues by comprehensively considering all aspects of drivers' personal, behavioral, academic and inclination details in dynamically classifying the drivers. The target classification mechanism categorizes the drivers' panoramically in four classes and retains the current driver's profile for future reuse. Lastly, this work is the first time in which CBR and Fuzzy logic have been applied in driver categorization to the best of our knowledge.

III. Proposed Approach for Driver Categorization

A modular view of the proposed approach with contemporary techniques is presented in Fig 1. There are four modules, namely, *Case base Reasoning* (CBR), *Artificial Neural Networks* (ANN), *Fuzzy Logic* (FL) and *D-CHAIT*. A comprehensive elaboration and representation of driver's attributes, preprocessing of these attributes for selecting most relevant ones, and machine learning techniques are furnished in following sections.

A. *Drivers Dataset*

The foremost aspect is to build the profile of the drivers. Three aspects play an important role while building a case base, i.e. format in which these cases are stored, attributes contained in a single tuple of the case base and the quality of data contained in tuples.

The performance of machine learning techniques greatly relies on the quality of the dataset used for training, so it is important to provide a glimpse of such dataset. All the implicit attributes were acquired from driver's institutes whereas explicit attributes were derived from drivers' input.

Driver's profile attributes modeled in Fig. 2. The highlighted part of features has minimal impact on classification of drivers as asserted by preprocessing in section 3.2.

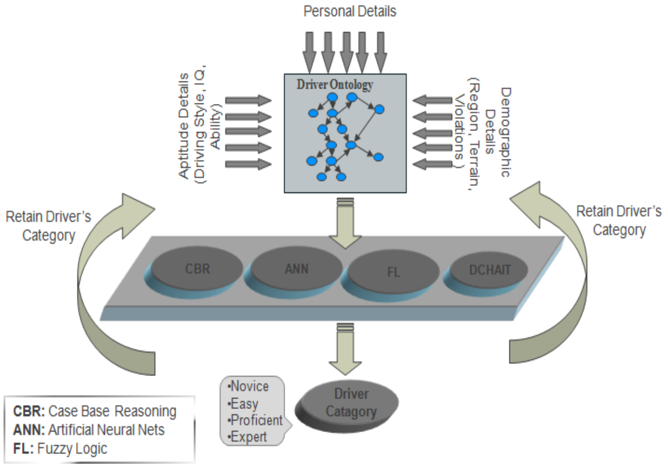

Fig 1. Proposed Approach for Driver Categorization (*D-CHAIT*)

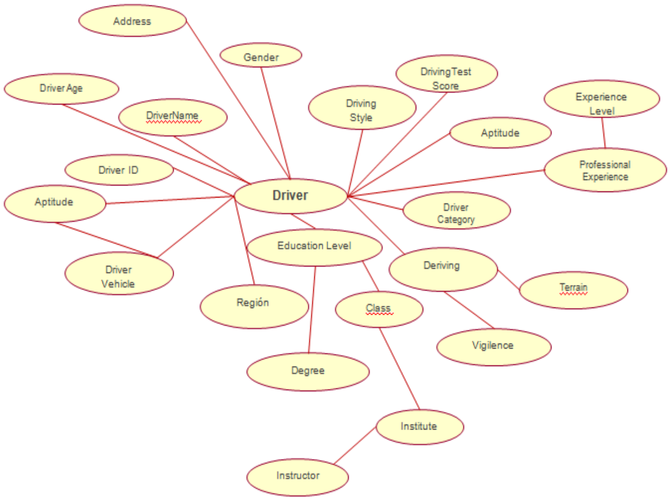

Fig 2. Driver Attributes, A Tuple in CBR's Case Base

In total, there were profiles of 800 drivers, each having 24 profile attributes for correctly categorizing the drivers. In other words, every tuple in the data set contained 24 attributes (columns), on the basis of which a driver was assigned a category. However, before subjecting this data set to machine learning techniques, it was preprocessed [21, 22] as explained in the next section. Different machine learning techniques addressing the issue of driver categorization so far have not considered the phase of preprocessing to the best of our knowledge.

B. *Preprocessing of Attributes and Data*

All the attributes in the data set of driver profiles were not expressive enough for playing a significant role in their categorization, may not have contributed towards classification accuracy and may have caused overfitting. Therefore, a preprocessing mechanism was employed to prune the most important attributes out of attributes selected initially.

TABLE 1
SELECTION OF FEATURE ATTRIBUTES

Types of Features	**Feature Set Presented: (Driver Category)**	**Feature Set Selected by WEKA**
Demographic	DriverID, Name, Age, Qualification, Region, Company	DriverID, Age, Region
Professional	RefresherCourse, DrivingTestScore, LicenseType, Institute, Award, InstructorFeedback violations, major accidents	DriverTestScore InstructorFeedback violations
Behavioural	DrivingStyle, Aptitude, Professional Capacity, Behavioural Rank, vigilence	Aptitude, Driving-Style
Inclination	Driving-Institute, Terrain, Experience Years, Vehicle Types, terrain	Experience Years, VehicleType, terrain

Preprocessing of these records was done given the metadata of attributes, the "Classes" in which the categorization was to be done and actual data (comprising of data values with their corresponding category) [23]. Features having a maximum impact on classifying a driver into a certain category are suggested by this phase. The number of features presented and the ones selected are given in Table 1.

The attributes which have been suggested after preprocessing of data are numerically represented on a scale of 1 to 10 except. This numeric representation offers twofold benefits: first, it is easy to measure/quantify the driver attributes and second, it can be easily fed to machine learning techniques as these techniques are tailored to operate on numeric data. Here it is worth mentioning that the attributes suggested in Table 1 have a strong impact on properly classifying the drivers.

C. *Case Base Reasoning based Classification*

Case based reasoning targets to resolve problems based on prior knowledge maintained in a case base. Whenever new drivers were enrolled in system, their profile was created by taking their personal details and ones pertinent to their aptitude and professional standing. Based upon this information, each driver was assigned a category reference to his profile strength i.e. *easy, novice, proficient and expert*. This category was maintained along with the rest of the profile details of the driver in a repository. This repository serves as a "Case base" for our CBR model that not only plays a key role in categorizing new drivers but is evolving over time.

Once the dataset is finalized after the preprocessing phase (Dataset serves as case base of CBR) with profiles of the drivers, a way to retrieve similar cases from the CBR's case base needs to be devised for finding out the similar drivers given the profile attributes of the new driver.

Case Retrieval: provides a query specific solution given the profile attributes of a new driver (query case). Level of similarity is computed for the *query case* against cases in the case base through the similarity metrics. A number of similarity mechanisms exist in the literature to find the similarity between a new case and a case in the case-base [24]. Different similarity metrics were considered for retrieval against *query case(s)* such as *Euclidean distance*, *Levnasthanian Edit distance* and *Taversky's Ratio Model* [25]. The *Euclidean distance* exactly compares each attribute of the new profile tuple with every corresponding attribute of the tuples in the case base and assigns a proximity rank to each of the matching tuples. On the other hand, the *Edit distance* calculates the cost of transforming the new case into every corresponding case of the CBR's case base. Addition, deletion and substitution operations are performed for this transformation. In our case, it does not maintain attribute sequencing and hence it was not feasible; therefore, *Taversky's Ratio Model* [26] was employed as a similarity metric due to its simplistic approach and degree of calculating *one-one* attribute similarity in every case.

If cases retrieved from case base appear with exact similarity i.e. driver attributes in the query case and the cases in the case base are the same, then the new driver is assigned the same category as that of a similar driver in the case base (termed as *Reuse* in CBR). The situation is quite straightforward till the point that we have similar cases retrieved from the case base (similarity threshold is kept at 70% after experimenting different thresholds). However, an *adaptation or revision* mechanism is desired to acquire a solution in the case of similarity being less than the specified threshold.

Case Revision aids in providing the possibly nearest solution by assigning a category to certain driver, if exact match for a new driver case is not found. A couple of techniques have been employed for case adaptation i.e. through '*Majority Vote Classifier (MVC)*' [26] and *Artificial Neural Networks (ANN)*. In MVC, occurrences of certain solutions are considered among the retrieved cases for classifying a certain driver. The driver category having a maximum number of occurrences is considered as the category of the new driver. In other words, the value of the *n*-th element is considered for selecting the most probable candidate. For example, if the case retrieval process returns 10 cases (each case corresponding to 10 drivers); 4 with category 'easy', 3 with category 'proficient', 2 with category 'novice' and 1 with category 'expert'; the category 'easy' is assigned to the new case (driver). The role of ANN in adaptation has been explained in the section addressing *LCHAIT* based classification.

D. Artificial Neural Networks based Classification

The Multilayer Perceptron (MLP) model of Artificial Neural Networks (ANN) [27] has been employed for driver categorization. The MLP has been selected due to its ability of regulating network weight in order to minimize the *Mean Square Error* (MSE). The MLP model was implemented using the *Neural Pattern Recognition* tool of Matlab 2015a with standard weights and activation functions. Besides, another script was written in Matlab separately for experimenting with different number of neurons and middle layers. The input layer contained 7 neurons, 2 hidden layers each with 8 neurons and an output layer with 1 neuron was used.

This ANN model has been trained over the same data contained in the case base of the CBR model as discussed in the section addressing CBR based classification. Moreover, the same set of query cases were used for testing the performance of the ANN model as in CBR. In order to train the ANN model, the dataset fed (i.e. the whole case base of CBR) was divided into three bins of training set, validation set and testing set. Training and validation phases were targeted for making adjustments to the ANN model in reference with its error rate and generalization. Subsequently, the testing phase measured the model performance with respect to its accuracy. Moreover, it aids in deciding if the ANN model needs to be retained provided that the error rate exceeds that expected. The dataset of our model was divided into three sets with a division of 70% for training, 20% for validation and 10% for testing of the ANN model. The performance of the model during validation/testing phases has been measured in terms of how accurately drivers have been classified while considering the associated costs.

E. Fuzzy Logic based Classification

Fuzzy logic can be considered as knowledge-based systems incorporating human knowledge into their knowledge base through fuzzy rules and fuzzy membership functions [28] by manipulating the linguistic data of driver. This module exploits the *Fuzzy Control Logic* in order to categorize the driver.

Whenever a new driver comes in, input variables (feature attributes selected) corresponding to driver's profile are fed into the Fuzzy logic model in crisp form scaled over a numeric range. For example, *PreTestScore* is an input variable with four ranges for fuzzification through a membership function i.e. poor (0-1.9), fair (2-4.9), good (5-7.9) and very good (8-10). These variables are fuzzified using the "Gaussian" membership function and represented in fig 3.

The Rule base of the fuzzy logic model aids in deciding the category of the driver. The knowledge required for the reasoning purpose is greatly dependent upon rules in the rule engine. Currently, there are 24 rules (*if-then-else*) in the current model of fuzzy inference engine. Some of these rules are presented next.

RULE 1: IF DrivingTestScore IS poor OR Qualification IS Inter OR DrivingStyle is Rash THEN DriverCategory IS Novice;
RULE 2: IF DrivingTestScore IS fair OR DrivingStyle is Stable THEN DriverCategory IS easy;
RULE 3: IF DrivingTestScore IS good AND Qualification IS Inter AND ViolationCount<3THEN DriverCategory IS proficient;
RULE 4: IF CGPA IS excellent AND LearningStyle IS good OR DrivingTestScore IS veryGood THEN DriverCategory IS expert;

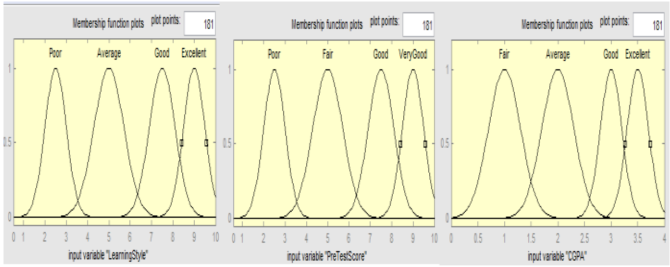

Fig 3. Membership function plots for Input Variables

The score assigned to each of the output attribute i.e. *DriverCategory* ranges from 0 to 10; for example driver's category for different drivers is 0 to 2.5 for 'novice', 2.6 to 5 for 'easy', 5.1 to 7.5 for 'proficient' and beyond 7.6 is 'expert'. A sample output variable with membership function is plotted in Fig 4.

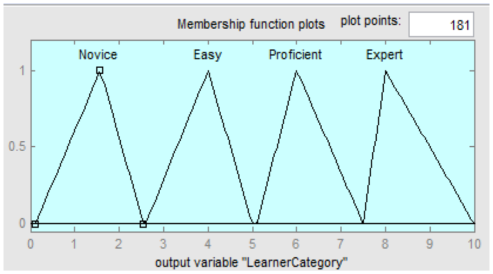

Fig 4. Membership Function Plot for Output Variable

After the rule engine yields a certain value for the driver, it needs to be transformed into human-understandable format i.e. defuzzification. The Centre of Gravity method has been used to defuzzify the output of the rule inference engine with other options of weighted average [29] and singleton methods.

F. D-CHAIT based Classification

The proposed technique *D-CHAIT* has been used to categorize the drivers with CBR and ANN. As the name suggests *D-CHAIT (Driver Categorization with Hybrid of Artificial Intelligence Techniques)* is a hybrid of two ML techniques i.e. CBR and ANN. For new drivers, profiles are retrieved in the same fashion as in the CBR retrieval phase using a certain similarity metric as discussed in section 3.3.

Among the retrieved profiles, cases that can be utilized are reused and the rest of them may be adapted for usage. The ANN model, used for categorization of drivers through CBR's adaptation phase, is trained over the retrieved cases instead of training on driver profiles contained in the whole case base as in section 3.4. Once the driver category is assigned to the drivers into any of the four categories by the ANN, the profile is retained into the CBR's case base for future re-use.

A mathematical representation of proposed technique based on extension of [30] is given in the following to formally describe the process of driver categorization:

Let (DA) represent an attribute of driver profile (DP) in the orm of a matrix with dimension of m x n:

$$\begin{bmatrix} (DP)_1 = (DA)_1, (DA)_2, (DA)_3, \text{--------} (DA)_n \\ (DP)_m = (DA)_1, (DA)_2, (DA)_3, \text{-------} (DA)_n \end{bmatrix}$$

where $((DP)_m \text{ x } (DA)_n) \epsilon \phi^{m \times n}$; here ϕ represents the case base of CBR, m=1400, n=9.

$$S(P,Q) = \{1 \text{ if } R \geq 0\} \vee \{0 \text{ otherwise}\} \qquad (1)$$

where

P = number of cases in the case memory,

Q= a query case,

T: threshold of similarity for reusing/revising a case,

R: is similarity rank of query case

Driver category can be predicted ('y+1') that is produced by 'y(n)' through driver-profile of '$\phi(n)$' in data set:

$$y(n+1) = f(\phi(n), y(n)) \qquad (2)$$

y(n+1) is the driver category predicted.

Input, state and output can be represented by:

$$y = A_x + B_y \qquad (3)$$

where $y \epsilon \phi^{maximum}$ with input $x \epsilon \phi^{maximum}$ (x^n) is state and $y \epsilon \phi^{maximum}$, (x^p) is the output.

The polynomial form of eq (2) can be given as:

$$\Phi(d/dt). Y = M(d/dt)\, x \text{ with } y = cal(x,y) \qquad (4)$$

CBR part is implemented in Java whereas the ANN part has been implemented in the same way as explained in section 4.2 with certain modifications in the MLP model. In order to retrieve the most relevant cases the same similarity metrics were used i.e. Tvesky's *Ratio Model* with the same similarity threshold as previously. Subsequently, the ANN model is trained on these retrieved cases only. This model of ANN showed optimal results with an input layer having 7 neurons, 1 hidden layer each having 8 neurons, and an output layer with 1 neuron.

The retrieved set of cases used for training the ANN model are divided into training set, validation set and testing set with proportion of 70% for training, 20% for validation and 10% for testing of the ANN model. The performance of the ANN model has been measured in the testing phase in terms of information retrieval metrics and associated cost as discussed in the next section.

IV. Evaluations and Discussion

In order to evaluate the effectiveness of the ML techniques and that of *D-CHAIT*, a dataset comprising of the profiles of 600 drivers was used. In order to build the profiles, data of drivers were acquired from different public sector driving and licensing institutes (NHA Rawalpindi[1], NMP Lahore[2], and CTP Peshawar[3]) for drivers, renewal cases of licenses, canceled licenses, international licenses for different age groups and regions. This data diversity was purposefully introduced to comprehensively cover a variety of cases in our data model. Comprehensiveness of data models ensures an effective training approach of the ML models independent of any biases (i.e. lack of coverage of cases or overfitting).

The input for the evaluation of the proposed techniques consisted in eight sets of new drivers' profiles (each set having profiles of 20 drivers). These 160 profiles were subjected as input to all the ML models of CBR, Fuzzy Logic, ANN and *DCHAIT* randomly for evaluating performance of ML techniques without to check coverage of ML model for every possible scenario.

However, for ANN out of 600 driver profiles, 300 were used as training and 140 were used for validation. Whereas, the same sets of new driver's profiles were used for testing the ANN model. The effectiveness of the ANN model was measured through its accuracy in assigning a category to the *drivers* presented as the validation set.

D-CHAIT has been evaluated through a variation of training set and validation set with same testing sets as in the scenarios above. Driver profiles retrieved through the CBR's retrieval phase have been used for training the ANN model (Retrieval mechanism of CBR is explained in the section discussing CBR based classification). The input for profile retrieval is given without provision of driver category, whereas the retrieved cases contain the driver' category (used for ANN training) while predicting the driver category.

Another dimension of our work is to categorize the drivers using Fuzzy logic. The same set of new driver profiles are fed to the fuzzy logic model for driver's categorization while exploiting the fuzzy inference engine.

The output of these models was recorded and evaluated through standard metrics in information retrieval to get a comprehensive picture of the performance shown by ML techniques.

Summary of accuracy in predicting degree of accuracy shown by different approaches is furnished in Table 2, where every value is the percent average depiction for performance models. It presents the average accuracy exhibited by different techniques in driver categorization. Figure 5 is simply a picture of accuracy in categorizing the drivers without taking into account aspects of precision and recall.

TABLE 2
COMPARATIVE ANALYSIS WITH RESPECT TO ACCURACY (AVERAGE)

Technique	**FL**	**CBR**	**ANN**	**D-CHAIT**
Average (%)	29.67	47.35	57.52	70.84

Literature [31] states that accuracy alone may not provide insight to the effectiveness of the ML based classifiers. So precision, recall and F-measure have also been used for representing the performance measures of all four techniques.

Average of standard precision (P) recall (R) and F-measure (F) are computed as presented Fig 6, 7 and Fig 8 respectively; provide a fine picture of performance analysis exhibited by techniques under experiment in terms of precision, recall and f-measure. It may be observed that CBR (52.87%, 61.87%, 55.87%) beats FL (32.50%, 44.37%, 37.52%), which in turn shows an inferior performance than ANN (59.12, 67.88, 62.33%) and *D-CHAIT*. Moreover, *D-CHAIT* (74.12%, 81.75%, and 78.33%) has better performance than rest of the prevalent techniques.

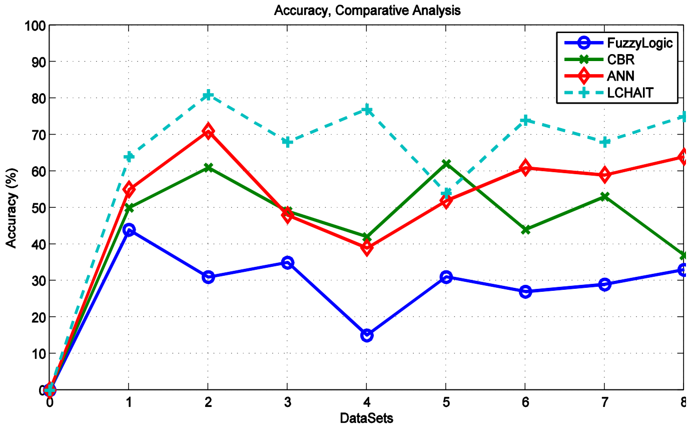

Fig 5. Comparative Analysis: Accuracy

Fuzzy logic, driven through the knowledge in the *Rule Inference Engine*, seems not adaptive to comprehend different scenarios with different parameter values due to its non-adaptive rule base.

For example, a driver with poor *test score* and average *education* is categorized as *Novice* according to rules but his good performance in *DrivingTest* suggests the categorization as *Easy*-level driver.

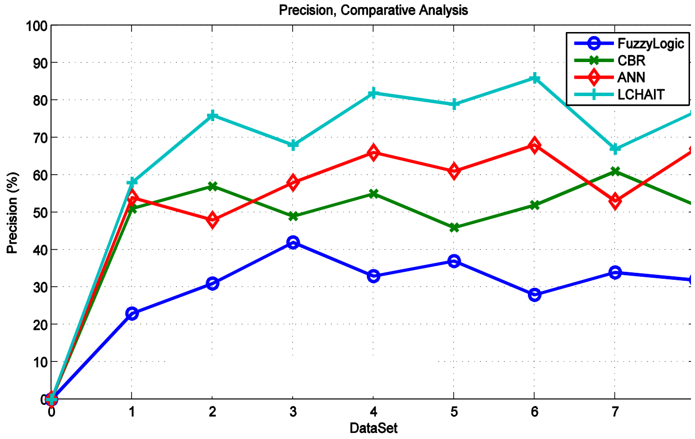

Fig 6: Comparative Analysis: Precision

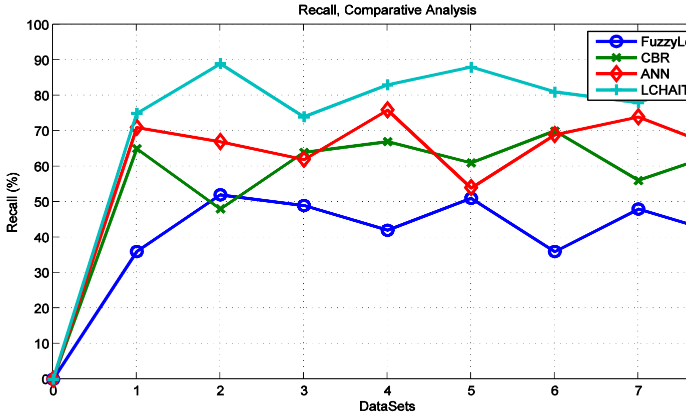

Fig 7: Comparative Analysis: Recall

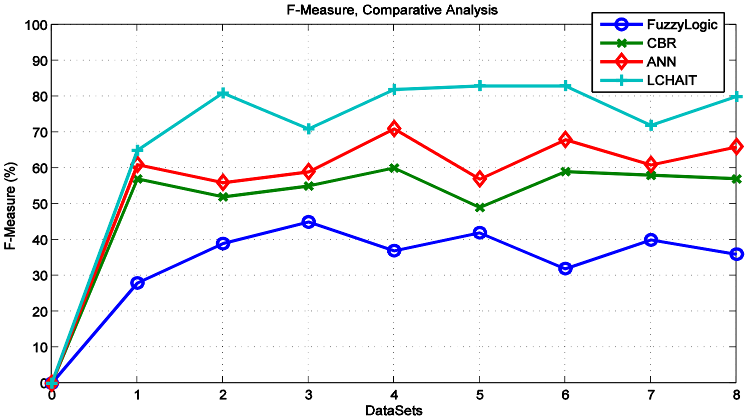

Fig 8. Comparative Analysis: F-Measure

Nevertheless *DrivingTest* has a higher impact in classifying the driver to certain class but rules in the fuzzy inference engine cannot comprehend these relationships. Such variation in profile attributes is not handled adaptively by the rule-base of fuzzy logic.

The performance of CBR depends upon the accurate selection of cases during retrieval phase through similarity metrics. Similarity metric (based upon ratio model) measures the one-to-one nearness of feature attributes of new driver and ones in the case base without taking into account the degree of similarity among features.

Secondly, relevant profiles are selected based on a static rank that inherently would ignore cases even if they are to be selected with a least margin without taking into account any exceptions. Thirdly, the adaptation phase with MVC itself suggests outputs in a static way by opting the value occurring maximum times among the retrieved profiles without considering relationship of attributes and classes dynamically. For example, a driver with good experience, average agility in driving style and average aptitude may be placed in an *Easy-level* category. But with these attribute values at borderline, categorization may falsely be done as *Novice* since the *Terrain* for which the category is assigned seems different. Such misclassifications eventuate due to the inadequacy of fragility in CBR and MCV.

ANN exhibits better performance than FL and CBR due to its dynamic and adaptive nature. Besides its input and output layers, there were two middle layers, each containing 15 neurons in order to get trained, validate and test the driver profiles for right categorization of new drivers.

On the other hand, *D-CHAIT* shows better performance than the rest of the contemporary techniques in terms of performance parameters so far. *D-CHAIT*, contrary to ANN, uses only one hidden layer with 8 neurons on the middle layer (computationally less expensive). The effectiveness of *D-CHAIT* is signified by the phase of preprocessing (both for attributes and data), by data used for training, and more importantly by training of the ANN model on the most relevant profiles retrieved through the CBR's profile retrieval phase.

However, ML techniques may not be assertive without taking into account the associated costs in terms of mean square error or percent error during the phases of training, validation and testing. These cost comparisons are not possible for CBR and FL due to their inherent nature. So costs incurred by ANN and *D-CHAIT* are compared in Fig 9a and 9b, respectively.

For an equal number of iterations, MSE costs for training, validation and testing of ANN reduces gradually with best estimate after comparing the actual and expected output. A spiky albeit minor behavior is observed for ANN with a bit of *overfitting/bias* in the training set. The *D-CHAIT* model, on the other side, performs better by revealing a smooth relationship between validations and testing curves (correct behavior with uniform and succinct training set) along a reasonable decrease in error rate.

Here, it is worth mentioning that different tools and technologies such as JAVA, Matlab, WEKA and Protégé have been used for implementing different modules of same framework that is proposed. However, the end to end developed system, as envisioned, would not have performance issues with respect to execution time with advent of advanced processing units [36].

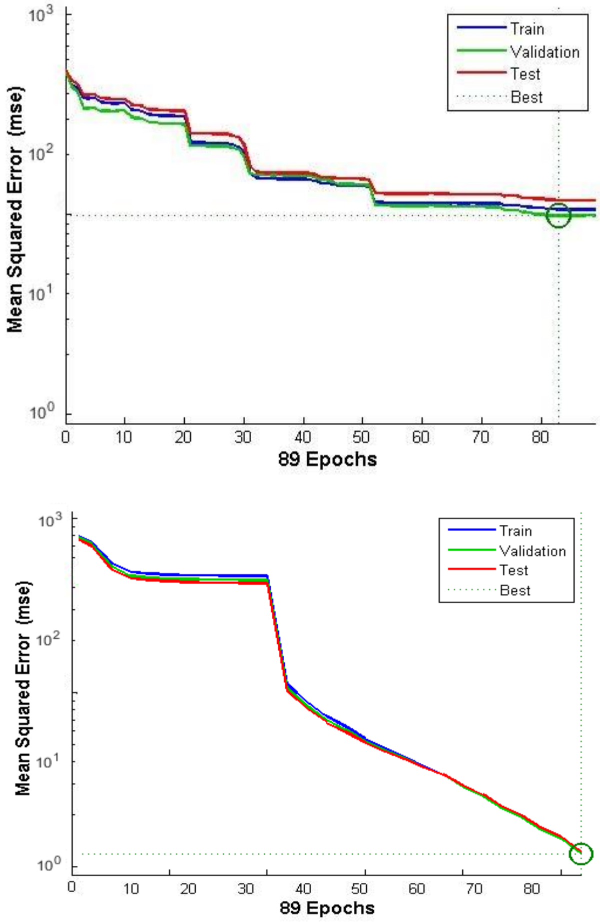

Fig 9. Comparative Analysis: (a) *ANN* (b) *D-CHAIT*

V. Conclusions and Future Work

Driver categorization targeted for vehicle systems is carried out through different ML techniques in this work. A comparative analysis for deciding the best one among Fuzzy Logic, Case Based Reasoning, and Artificial Neural Networks techniques is presented followed by proposing a novel technique named *D-CHAIT*, a hybrid of ML techniques. A rigorous empirical analysis, based on different evaluation metrics, suggests the premise of *D-CHAIT* as a better choice for driver categorization due to its dynamic and adaptive nature. The effectiveness of the approach is not only courtesy to preprocessing, data cleansing and effective training set but also to the inherent features of CBR retrieval and adaptivity instilled in by ANN.

The CBR module of *DCHAIT* uses similarity metrics in retrieving the relevant cases from the case base. The techniques used seem trivial and static. So, different similarity metrics such as clustering or fuzzy logic would be employed to experiment unsupervised and supervised techniques for dynamic retrieval of relevant cases. For the adaptation part, we look forward to experiment with another variation of neural networks as given in [32] that works well with limited training/validation sets. Another dimension may be to experiment with fuzzy logic by making its rule base dynamic through Genetic Algorithms (GA) as done by [33].

We look forward to employ the *D-CHAIT* model in a real time semantic driver categorization system for the categorization of drivers. The envisaged system would recommend vehicles with varying levels of terrains to the supervisory managers. Moreover, the driver category would be made dynamic to take into account the re-categorization with reference to the performance of the drivers in a certain span of time.

The future of automated vehicles has been attributed to a changed role of derivers [38] that would raise safety concerns, confusion and traffic conflicts. Moreover, the research challenge for having a partial but effective role of drivers will persist. So, proposed framework is envisaged to cater emerging challenges in domain of vehicular technologies.

Acknowledgment

Authors of this manuscript are thankful for the support of Ontology Engineering Group (OEG) in unconditional support and guidance for completing this research. Also, we wish to acknowledge our gratitude to Higher Education Commission of Pakistan in extending resources for completing this work.

References

[1]. P. Angkititrakul, J. Hansen, “UTDrive: The smart Vehicle Project”, *the 2007 Biennial on DSP for In-Vehicle and Mobile Systems*, 2007.

[2]. A. Baron, P. Green, “Safety and Usability of Speech Interfaces for In-Vehicle Tasks while Driving: A Brief Literature Review”, *UMTRI-2006-5: The University of Michigan, Transportation Research Institute*, pp.1-8, 2006.

[3]. C. Bishop, Neural Networks for Pattern Recognition, Oxford: Clarendon Press, 1995

[4]. Sangjo Choi, Jeog Kim, John Hansen, "Analysis and Classification of Driver Behavior Using in-Vehicle CAN-BUS Information", *INTERSPEECH*, 2007

[5]. G. Albert, O. Musicant, T. Toledo, "Evaluating Changes in the Driving Behavior of Young Drivers a Few Years After Licensure Using In-Vehicle Data Recorders", 2012

[6]. Y. Chen, H. Peng, "Modelling of uncertain reactive human driving behavior: a classification approach", *IEEE Transactions on Control Systems Technology*, 1-13, 2013

[7]. M. He , I. Baek "Vehicle's Lane-changing Behavior Detection", http://mi.eng.cam.ac.uk/projects/segnet/., 2012

[8]. D. Salvucci. "Inferring driver intent: A case study in lane-change detection", *Human Factors and Ergonomics Society Annual Meeting*, vol. 48, no. 1, pp. 2228–2231, 2004.

[9]. Sankar K., Simon C., SHIU "Foundation Of Soft Case based Reasoning" by John Wiley & Sons, Inc. 2004.

[10]. D. Arditi, B. Tokdemir, "Comparison of Case-Based Reasoning and Artificial INeural Networks"; Journal of Computing in civil engineering, July 1999.

[11]. Ying Bai and Dali, "Wang Fundamentals of Fuzzy Logic Control – Fuzzy Sets, Fuzzy Rules and Defuzzifications", Advanced Fuzzy Logic Technologies in Industrial Applications. Springer 2006.

[12]. Romero, S. Ventura, "Educational Data Mining: A Survey from 1995 to 2005", *Expert Systems with Applications*, vol. 33, no. 1, pp. 135-146, 2007.

[13]. P. Kaur, "Classification and prediction based on DM algorithm for slow learners". *International Conference on Recent Trends in Computing* 2015.

[14]. M. Cristian, "Classification of Learners Using Linear Regression", *Federated Conference on Computer Science and Information Systems* pp. 717–721, 2011.

[15]. G. Chen, C. Liu, "Discovering decision knowledge from web log portfolio for managing classroom processes by applying decision tree and data cube technology", *Journal of E-Computing Research*, vol. 23, no. 3, pp. 305–332, 2000.

[16]. J. Han, M. Kamber, "Data Mining-Concepts and Techniques", Morgan Kaufmann Publishers. 2011.

[17]. B. Saleena., K. Srivastava, "Using concept similarity in cross ontology for adaptive e-learning", *Journal of King Saud University- Computer and Information Sciences*, vol. 27, no. 1, pp. 1-12, 2015

[18]. E. Frank, A. Mark, H. Ian, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

[19]. E. Pothos , J. Trueblood, "Structured representations in a quantum probability model of similarity", *Journal of Mathematical Psychology*, vol. 64, no. 5, pp.35-43, 2015.

[20]. X. Chang, H. Li, J. Rong, Z. Huang, X. Chen, and Y. Zhang, Effects of on-Board Unit on Driving Behavior in Connected Vehicle Traffic Flow, Journal of Advanced Transportation, vol. 2019, no. 3, pp 1-12, Article ID 8591623, 2019.

[21]. Bishop, Christopher, "Pattern Recognition and Machine Learning", Springer-Verlag, 2006

[22]. E. Ukkonen "Constructing Suffix Trees On-Line in Linear Time". In Algorithms, Software, Architecture", *12th World Computer Congress, Madrid*, Spain, Elsevier Sci. pp. 484-492, 1992.

[23]. C. J. Keith, Van Rijsbergen, "A new theoretical framework for information retrieval" *9th annual international ACM SIGIR conference on Research and development in information retrieval* pp. 194 - 200, 1986.

[24]. G. Finnie, Z. Sunt, "Similarity and metrics in case-based reasoning", *International Journal of Intelligent Systems*, vol. 17, no. 41, 273–287, 2002.

[25]. P. Cunningham, A taxonomy of similarity mechanisms for case-based reasoning. *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 8, pp. 1532–1543, 2009.

[26]. L. Lam, Y. Ching, "Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance", *IEEE Transactions on systems, man, and cybernetics Part A: Systems and Humans*, vol. 27, no. 5, 1997.

[27]. M. Constantin, E. Valentina, "Multilayer Perceptron and Neural Networks", *Wseas Transactions On Circuits And Systems*, vol. 7, no. 8, pp. 2-10, 2009

[28]. D. Srinavasan, L. Cheu, "Hybrid Fuzzy Logic-Genetic Algorithm technique for automated detection of traffic incidents on freeways", *IEEE Intelligent Transportation Systems Conference Proceedings*, pp. 7194-7201, 2001.

[29]. S.M. Odeh and et al, "A hybrid fuzzy genetic algorithm for adaptive traffic signal system", *Advances in Fuzzy Systems*, pp. 1-11, 2015.

[30]. S. Sarwar, Z. Qayyum, R. Castro, M. Safyan, R. Munir, M. Iqbal, Ontology based E-learning Framework: A Personalized, Adaptive and Context Aware Model Multimedia Tools and Applications, vol 77 (3), pp. 26-39, 2019

[31]. K. Holger, F. Ray W, N. Natalya, "The Protégé OWL plug-in: an open development environment for semantic web applications", *The semantic web ISWC 2004*. Springer, pp 229-243, 2004.

[32]. S. Beom, S. Kwun, "Design of Radial Basis Function Neural Networks with Principal Component Analysis and Linear Discriminant Analysis for Black Plastic Identification", 17th ISIS, 2016.

[33]. M. Bhaskar, M. Das, "Genetic Algorithm Based Adaptive Learning Scheme Generation for Context Aware E-Learning", International *Journal on Computer Science and Engineering* vol. 2. no. 4, pp. 1271-1279, 2010.

[34]. C. Oscar, F. Mariano, G. Asunción. "Ontological engineering: what are ontologies and how can we build them?", Idea Group Inc.; 2007.

[35]. F. Lopez, G. Perez, N. Juristo, “Methontology: from ontological art towards ontological engineering”, *Symposium on ontological engineering of AAAI*; 1997.

[36]. http://girtab.dia.fi.upm.es/webAR2DTool/, online tool for ontology representation developed at Ontology Engineering Group (OEG), UPM Spain. Accessed on 19 April, 2016.

[37]. C. Gatenberg, Intel announces its latest 9th Gen chips, including its ‘best gaming processor’ Core i9 (28 Core Xeon processor), 2018

[38]. I. Noya David, S.William J.Horreya, Automated Driving: Safety Blind spots, Elsevier Journal of Safety Science, vol. 102, no 6, pp. 68-78, 2018.