Deep learning Assisted Predictive Analytics

Juby Mathew

Abstract: Big Data has become a dominant term in describing the exponential growth, accessibility, availability, and widespread use of information—in structured, semi-structured and unstructured format—in a variety of business context. Of the many areas for application, Big Data Analytics is capable of making a major mark in road safety. Analysis of road accidents, vehicular crashes and inflicted casualties and damage are significant areas of application. Many researchers have tried to solve these issues but still, there are gaps in the road accident severity prediction and finding the contributory factors such as season and time of the accident in which the accident frequently occurred. This leads to the challenges in the field of accident analysis and prediction. The investigation of the risk factors that contribute to the severity of the injury in motorcycle crashes is provoking as a challenge problem across the globe. With its intention, a novel prediction model that predicts the risk factors contributing injury severity in motor vehicle crashes is introduced in this work. Initially, the risk factors contributing injury severity like the accident information, vehicle information, personal information and other information are collected. Further, the prediction process is handled using deep learning model, where optimized Deep Convolutional Neural Network (Deep CNN) is used. In order to enhance the performance of prediction, certain parameters of the convolutional layer, dense layer and dropout layer are fine-tuned by a new Sea Lion updated Dragonfly Algorithm (SL-DA) model, which is the hybridized version of Sea Lion Optimization Algorithm (SLO) and Dragonfly Algorithm (DA). Further, the performance of proposed work will be compared over other state-of-the-art models with respect to positive and negative measures as well.

Keywords: Big Data, Prediction model, Deep Learning, Feature extraction, Automobile Crashes, Injury severity, Risk factors

I. INTRODUCTION

Nowadays, big data is the most admirable research topic, however, the mining of needed data from a huge volume of data is considered as the major issue. Number of researches are in progress to find the solution for this problem. The solutions often vary from some conventional data mining approaches, by which the process of mining should be more effective and improvable. Handling of big data offers an issue to conventional computation daises and hardware as well. As we, all know big data is a concept when the data becomes so large and could not be handled by the traditional database systems and other conventional data handling concepts. There are many methods for data analytics in big data. As deep learning process provides much benefits than the traditional big data analytic methods. This paper focus entirely on utilizing the deep learning technics in analytics of big data. Many methods are deployed in the analytics of data to understand the data structure and the relationship between them with respect to Big data analytics. These models are conventional statistical models and there are many tools available in the market to do data analysis namely, apache Hadoop, apache spark, etc.

The Road Traffic Accidents (RTA) is rising as a largest national health issues in the recent years and this is due to the diverse factors like driver, environment, car, etc. Few among these parameters are more crucial for determining the severity of the injury. The intensity of the accident is affected by several attributes like the "road conditions, weather conditions etc". The major root cause for these traffic accidents and the injury severity related to the crashes are of exceptional concern to the public, but to the academic as well as industry and government. Since the evaluations by these industries not only aims at preventing the crashes, but also tries to lessen the severity of the outcomes by saving money and life. Hence it is more crucial to evaluate of the accident [4] [5]. It is more crucial to these accidents to identify the features that lead to the evaluate these accidents in order to identify the features that lead to the criticality of the accident. The major reason behind the prediction of the injury severity in the accidents is to provide appropriate medical assistance in time to the one, who is involved in accident. This can reduce the casualties in the accident, inform the corresponding emergency decision making department in time, and avoid greater property losses.

Thus, data mining analysis provides a solution to this significantly increasing issue in order to identify as well as forecast the influential factors like man, automobile and environmental factors, and thus to explain severity of RTAs. Even though long runs reports are being provided by the organization against data warehouses, most of the organizations haven't opened these repositories to in-depth on demand exploration. This is mainly due to the complications taking place with the lack of analytics tools and due to the absence of the important data needed for the powerful evaluations.

II. DEEP LEARNING MODELS FOR PREDICTIVE ANALYTICS

A key property of Deep Learning is its ability to provide hierarchical representations for either labelled structured input data or unlabeled unstructured input data [1]. Hierarchical representation concentrates on learning high order representations from low-level data [5]. Recognizing words from audio, recognizing objects from images, and recognizing poses and movement from the video are considered as examples for this scenario. This property is the reason for the movement to Deep Learning in the computational statistics and machine learning community. The choice of deep learning models for predictive analytics is mainly due the fact that they can derive conclusion from a huge pile of data without much intricate engineering techniques. Also the same models are capable of dealing with structured and unstructured sequential data. Bengio et al.

Revised Manuscript Received on April 17, 2020.

* Correspondence Author
Dr. Juby Mathew*, Department of Computer Applications, Amal Jyothi College of Engineering, Kanjirapally, Kottayam, Kerala, India. Email: jubymathew@amaljyothi.ac.in
[16] introduced a model that learns word vector representations as part of a simple neural network architecture for language modelling.

A. How Predictive Analytics Work:

A precise and actual predictive analytics takes some upfront effort to set up. Predictive analytics needs experts who know that there must be a business problem to be resolved, models that require to be made and distinguished, and management to set the estimates into the act for optimistic results.

**Fig 1: Life Cycle of predictive modeling**

B. Predictive Modeling:

Predictive modeling is a commonly used statistical and data mining technique that works by analyzing historical and current data and generating a model to help predict future outcomes. It gives us the power to discover hidden relationships in volumes of data and use those insights to confidently predict the outcome of future events and interactions. Predictive modeling is normally used in an arithmetical method to guess future behavior. The predictive model works with the data sourced from various points. This follows the process of data transformation, which includes cleaning and processing of data, to a pattern easily comprehensible for analysis. In 2017, Vijayakumar et al. [17] developed Predictive model for classification using SVM. The figure 1 shows the life cycle of predictive modelling. Road traffic crashes are researched as one of the biggest public health and injury prevention problems. The issue is having wide reaching effects as the crash victims fall into a serious health problem suddenly from a state of good health. According to the WHO report (2009) an estimated 1.3 million people perish each year due to road accidents, around 20 to 50 million sustain non-fatal injuries. As of now, traffic accidents are the leading cause of death among children of 10 – 19 years of age [17]. Traffic accidents are detrimental to the safety of human life and property. Accidents are a result of indirect and direct influences of many factors. The solution to accident management is still a widely-researched area among researchers and a clear solution is still a distant dream [10]. Accident prediction models (APMs) have a multi-faceted role in their applications. They are capable of predicting the rate of accidents in different roadway entities such as highways, intersections, interstate boundaries, etc. These models can also pick out geometric, environmental and operational parameters that lead to an accident. Hence the need to scrutinize the relations that exist between the inter-related roadway entities is of utmost importance. [11].

**MOTIVATION**

The ultimate goal of accident analysis is to improve an agency’s ability to make future decisions in all components of a highway safety plan. There are several road traffic analysis and prediction models that are capable to prevent, manage and unravel the reasons behind traffic congestions. The forecasting or prediction of road traffic characteristics, such as speed, flow and occupancy, allows planning new road networks, modifications to existing road networks, or developing new traffic control strategies. This paper surveyed the latest studies in the field of traffic accident prediction. Due to the importance of the subject, the use of techniques that assist in the process of predicting traffic accidents and determining the most important factors has various impact on accidents. Listed below are the different techniques used in this study to predict traffic accidents:

**The major contribution of this work is as follows:**

- Introduces a new Optimization assisted deep learning-based prediction model is introduced that predicts the risk factors contributing injury severity in motor vehicle crashes
- The fine-tuning of CNN is done by means of a new hybrid algorithm termed Sea Lion updated Dragonfly Algorithm (SL-DA) approach
- The performance of proposed algorithm is compared and proved over other conventional methods with respect to certain performance measure.

**III. LITERATURE REVIEW**

In 2019, Zheng et al. [1] have proposed a novel TASP CNN model for predicting the harshness of the traffic accidents. This technique was features acquired from the combination relationships of the traffic accidents. They have constructed the FM2GI on the basis of the weight of the features acquired from the traffic accidents. The single feature relationship corresponding to the data of the traffic accident were converted into gray images having “combination relationships in parallel as the input variables for the model.” In 2019, Liu et al. [2] have developed an integrated spatio-temporal to disintegrate the gone down injury correlates that varies in the time and space domain. They have explored the spatio-temporal patterns associated with the pedestrian injury severity correlates in traffic crashes using the time and space-referenced crash data in motor vehicle crashes, the harshness of the perambulator damage was examined by means of deploying the GTWOLR Further, the results exhibited theta increased pedestrian injury severities are associated with crashes that involved pedestrian intoxication, SUV, heavy-duty vehicle (bus or truck), or that occurred after sunset without streetlights

In 2015, Olabarria et al. [3] developed a Bayesian classifier to classify traffic patterns as leading or not leading to collision and used 5-min standard deviation of speed as the indicator. In a later study, the same authors [3] refined the model and proposed a method to determine observation time slice duration prior to crash occurrence. Abdel-Aty et al. developed a matched case-control model to model the crash potential based on traffic loop data and the rain index.
In 2019, Juby et al.[4] found that the C-means clustering method, the Naive-Bayes method and the Discriminant Analysis was only able to identify the patterns leading to collision with an overall classification error rate at about 50%, and they analyzed reasons for the unsuccessful identification of the patterns leading to collision. In 2018, Jeong et al. [5] have developed a novel approach for classifying the severity of the injury in the motor vehicle crashes with the aid of “high accuracy and sensitivity rates.” The dataset for evaluation was acquired from MTCF dataset. Further, they have utilized the several techniques like the over-sampling and under-sampling for imbalanced classes. Further the level of severity of injury was classified by utilizing five classification learning models "Logistic regression, Decision tree, Neural network, Gradient boosting model, and Naïve Bayes classifier”. They have deployed the "Bootstrap aggregation (or bagging) and majority voting to chance the classification performance. The classification performance was evaluated using the geometric mean (G mean)

In 2019, Nishimoto et al. [6] have proffered a new serious injury risk prediction algorithm for pedestrians” with the aid of the data collected from the South Australian Traffic Accident Reporting System”. The proposed algorithm was based on the pedestrian crash data regression analysis. Further, an optimal model was developed with respect to the best combination of risk factors as defined by AIC. The model was developed on the basis of GPS data relating to the site and environmental conditions as well as pedestrian age gender, driver age gender and vehicle model year. This model had enhanced the pedestrian survival rate and had reduced the driver age gender and vehicle model year. This model had enhanced the pedestrian survival rate and had reduced the count of fatalities ultimately. In 2019, Zong et al. [7] have developed a novel severity causation network with the intention of exploring the risk relationship existing between the risk factors and the crash severity. This was accomplished by means of integrating the “information entropy with Bayesian network”. On the basis of the accessibility and weight, the influence of the different risk factors over the severity indexes was estimated quantitatively using the entropy weight method.

A. Review

The ultimate goal behind the accident injury severity prediction is to enhance the decision-making ability of the agency in all the highway safety plan components. A vast count of researchers in the area of the injury severity prediction was based on the crash related datasets. Although, most of them concentrated on traffic accident records limited to a small specific geographic region, a particular crash type, or a specific road or environmental conditions, they have paved the road for more comprehensive analytics studies. The most important works undergone under this subject is discussed in the literature section and a short of it is presented in Table I.

TASP-CNN in [1] is flexible and efficient in predicting the severity of the risk even under the worst cases. But, data sets of a traffic accident are imbalanced and hence lead to over-fitting. In addition, there exists no correlation between the features of the traffic accident’s data sets. Further, GTWOLR [2] take into consideration the temporal as well as the spatial nature of the traffic clashes. The time consumed for prediction is lower and hence not utilized in large scale. Naïve Bayes classifier in [5] has the potential of handling the imbalanced structure of the dataset. Apart from these advantages, it is less robust and more prone to classification errors. The AIC based approach in [6] is good in improving the pedestrian survival rate and reduces the number of fatalities used under a ‘worst-case scenario’. This technique is limited due to the utilization of smaller datasets. Moreover, information entropy with The learning vector quantization neural network model in [7] is good in enhancing the survival rate with its higher response time. But, this technique is not as accurate as the optimal model. Thus, all these most interesting works together contributed to developing a new injured severity prediction model with reduced errors.

### TABLE I. Features and Challenges of the existing accident severity prediction approaches

<table>
<thead>
<tr>
<th>Author (citation)</th>
<th>Methodology</th>
<th>Features</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zheng et al[1]</td>
<td>TASP-CN N</td>
<td>Availability, flexibility and high efficiency, Consumes less times</td>
<td>× Data imbalance × Prone to over-fitting × Requires high computational cost</td>
</tr>
<tr>
<td>Liu et al [2].</td>
<td>GTWOLR</td>
<td>Considers both spatial and temporal nature of traffic crashes</td>
<td>× Consumes more time</td>
</tr>
<tr>
<td>Jeong et al[5]</td>
<td>Naive Bayes classifier</td>
<td>Reduce the variance of out-of-sample error, Avoids over fitting</td>
<td>× Computation time is high × Less robust with high variance</td>
</tr>
<tr>
<td>Zong et al[7]</td>
<td>Information entropy with Bayesian network</td>
<td>Computes the actual probability of severity level, Improve the feasibility and efficiency of severity prediction</td>
<td>× Reduction of prediction accuracy × Unevenness of the dataset</td>
</tr>
</tbody>
</table>

IV. PROPOSED PREDICTION APPROACH

This paper intends to propose a new data analytics model under the automobile application. In fact, the model goes with the introduction of prediction model that predicts the risk factors contributing injury severity in motor vehicle crashes. The proposing prediction model involves phases like feature extraction and prediction. In the feature extraction phase, features related to accident information, vehicle information, personal information and other information are extracted. Further, the prediction process will be handled using deep learning model, where Deep Convolutional Neural Network (Deep CNN) will be used. Moreover, in order to enhance the performance of prediction, it is planned
to make the fine-tuning of activation function and count of convolutional layer by introducing a new Hybrid algorithm. The proposed algorithm is the combination of Sea Lion Optimization Algorithm (SLnO) and Dragonfly Algorithm (DA).[8] Both SLnO and DA[9] is the renowned optimization algorithm for resolving complex optimization crisis in different applications.

The injury severity risk factors in motor vehicle crashes are predicted by designing a novel prediction framework. The architecture of the proposed framework is manifest in Fig. 2. Initially, the accident information, vehicle information, personal information, and other information are extracted. This extracted information is subjected to the prediction process via optimized Deep CNN. As a novelty, certain parameters associated with the convolutional layer, dense layer, and dropout layers are optimized via a new hybrid optimization algorithm referred to SL-DA, which is the combination of SLnO and DA. Thus, the enhanced version makes the prediction more accurate and precise.

Fig 2. Proposed injury severity prediction model

V. SIMULATION SEVERITY PREDICTION MODEL

The proposed injury severity prediction model using the optimization deep CNN was implemented in Python platform and the corresponding outcomes acquired are noted. The database was from https://www.kaggle.com/c/accident-severity/data. Further, the presented work is compared over the existing works (both traditional and deep learning-based models) in terms of positive (like accuracy, precision, sensitivity, and specificity) and negative measures (FPR, FNR, and FDR). The traditional machine learning models utilized for evaluations are DT [12], KNN [13], RF [14] and SVM [15], respectively. In addition, the presented work (SL-DA+DCNN) is compared over the deep learning model like traditional CNN [17].

VI. PERFORMANCE EVALUATION

There a number of ways to evaluate the performance of DL methods. The recall, precision, and F1-score are a noteworthy performance measure for prediction, ranking and classification-based tasks. The F1-score stabilizes the recall and precision as:

\[ F1 = \frac{(Precision \times recall)}{(Precision + recall)} \]

where precision - the portion of pertinent instances among instances retrieved while recall - the ratio of retrieved pertinent instances with the total number of pertinent instances in the data [4].

VII. EXPECTED OUTCOME

The proposed model will be implemented in Python. Further, the performance of proposed work will be compared over other state-of-the-art models with respect to Type I and Type II measures. Here, Type I measures are positive measures like Accuracy, Sensitivity, Specificity, Precision, Negative Predictive Value (NPV), F1Score and Mathews correlation coefficient (MCC), and Type II measures are negative measures like False positive rate (FPR), False negative rate (FNR), and False Discovery Rate (FDR).

VIII. CONCLUSION

High dimensional and sparse nature of large data sets can pose a significant challenge in the accuracy of predictive modelling every day, the real world collects massive amounts of unlabeled real-time data coming from multiple resources; these data are referred by the term 'Big Data'. Deep Learning is a promising method to explore the details found within the system. The proliferation of high dimensional big data from various data sources becomes the main attention for data scientists. Thousands of people die in traffic crashes yearly. People lose their lives every day and more people are injured every hour. The present review has addressed different techniques that are deployed for traffic congestion and traffic crash prediction. Hence there is a need to address this gap in literature and provide a valuable contribution to this field. This paper had developed a novel prediction model that predicts the risk factors contributing injury severity in motor vehicle crashes, prediction model that predicts the risk factors contributing injury severity in motor vehicle crashes were gathered. Subsequently, they were subjected to optimized Deep Convolutional Neural Network (Deep CNN) for prediction process. As a novelty, certain parameters of the convolutional layer, dense layer and dropout layer were fine-tuned by a SL-DA algorithm, such that the prediction performance was increased. The proposed SL-DA algorithm was the hybridized version of traditional SLnO and DA.

REFERENCES

2. Jun Liu, Alexander Hainen, Xiaobing Li, Qifan Nie, Shashi Nambisan, “Pedestrian injury severity in motor vehicle crashes: An integrated spatio-temporal modeling approach”, Accident Analysis & Prevention, vol.132,November 2019
6. Tetsuya Nishimoto, Kazuhiro Kubota, Giulio Ponte,"A pedestrian serious injury risk prediction method based on posted speed limit" ,Accident Analysis & Prevention, vol.129,pp.84-93,August 2019

AUTHORS PROFILE

Dr Juby Mathew is a Dynamic, Resourceful Teaching Professional. He received his PhD and PostDoc in Computer Science from Mahatma Gandhi University, Kottayam. He pursued his MCA from Periyar University, Salem, MPhil in Computer Science from Madurai Kamaraj University and M.Tech from MS University, Tirunelveli. So far he has published his articles in 12 international Journals and presented papers in more than twenty National and International Conferences. Over 16 years of diversified teaching and corporate experience made him actively involved in all areas of education, including Curriculum Development, Student Mentoring, Student Career Preparation, and Community Work. At present, he is working as an Associate Professor in the Department of Computer Applications at Amal Jyothi College of Engineering, Kanjirapally, Kerala. He won Best Faculty award as a result of his enhanced ability to improve students performance, promising to shape a better world for the students and empower them with knowledge. He has reviewed many paper publications and journals and PhD thesis within an incredibly short period.