Road IQ: A Data based approach to understand and model Road Accident Risk Prediction

Anuj Rajeeva Singh (5305926) A.R.Singh@student.tudelft.nl

1 ABSTRACT

About 1.35 million people die each year as a result of road traffic crashes and an additional 30-40 million are severely injured or disabled. Not only do road accidents cause human loss, but are also a major cause for economic damage in every developed as well as developing country. Road traffic safety officials can facilitate traffic safety through various means such as law enforcement, education and awareness about the topic and by controlling traffic at specific high risk and crowded areas. But the fore-mentioned statistics are testimony to the fact that the current methods of enforcing transportation safety are not the most efficient and that there is still a huge scope for improvement. This paper aims at designing a system to predict the risk or probability of traffic accidents based on various spatio-temporal factors such as a detailed timestamp indicating the time and day of the week, coordinates of various regions in the geographic domain of research, clubbed with detailed road traffic information such as Annual Average Daily Traffic (AADT) for various vehicle types, population densities of regions, weather conditions, road types and various other informative dimensions. Machine learning algorithms that learn from historical data can alert users about the risk of an accident by fetching useful user data in real time and thus avoid the dangerous areas for that time. It is also useful in the context of better urban planning to create efficient systems of road maintenance and traffic management particularly with the upcoming advent of automated cars. This research also suggests how the satellite images of regions and roads scraped using Google Maps Static API can be used to enhance the predictive power of the structured data with the help of Deep Learning methods such as Convolutional Neural Networks (CNN). The data used for all the analysis and modelling has been obtained from the Department of Transport of Great Britain.

2 INTRODUCTION

Road traffic planning and organisation poses an crucial onus on the governments and city planners in the light of the grim road traffic accident statistics of the world. More than 1.2 million people are killed worldwide in road accidents, each year, the majority of which are aged between 15 and 29. Despite notable advancements in vehicle manufacturing, design and technology and road engineering, road traffic accidents remain as one of the leading causes of deaths and injuries. Apart from the massive damage to social welfare, traffic accidents cause countries up to 3% of their GDP([2]). Hence, it becomes imperative that more efficient methods must be developed in order to mitigate such losses at the worldwide scale. Poor infrastructure and maintenance of roads and services are the leading causes of road accidents and thus it would be safe to conclude that road accidents and deaths are more prevalent in low-income developing countries. Keeping this in mind, it is also important that the solutions being developed are cost effective and easy to use and deploy. This paper aims to understand and analyse the causes behind road traffic accidents and then develop an ensemble of Machine Learning and Deep Learning based algorithms that accurately suggests a *Risk Score* indicating the risk of an accident based on a number of factors. These factors have been obtained from the open data source of Great Britain - Road Safety Data, Road Traffic Counts data ([20]) and the Census Data of 2011. Data is easy to obtain and modelling it for personal use is cost effective too. However, certain specialised data generation requires setting up sensors throughout the road and highway network of the entire country. This may require a heavy budget and may also have certain shortcomings due to uncertain environmental factors. To this end, it is suitable to use satellite imagery of road and highways for the modelling and prediction tasks, obtained from the cloud based service of Google called Maps Static API. Using data to obtain a risk score can be used to alert the users in real time about the potential occurrence of an accident based on the continuous data being fetched. This not only mitigates the danger of an accident by alerting the users before the mishap and indicating the potential major factors that may cause such an event, it also could help in traffic planning. Assigning a score to every location/ coordinate on the road and highway networks opens up possibilities to use Multi-Agent Reinforcement Learning to plan and develop traffic planning strategies that aim to reduce traffic clogging and accident risk.

3 RELATED WORK

Related literature in this field focuses on pointing towards the factors that facilitate and increase the likelihood of accidents. Abdel-Aty et. al. [3] do the same by using Negative Binomial modeling techniques to model the frequency of accident occurrence and involvement, and results showed that heavy traffic volume, speeding, narrow lane width, larger number of lanes, urban roadway sections, narrow shoulder width and reduced median width increase the likelihood for accident involvement. Najjar et. al. [18] propose a novel method to use satellite imagery obtained from over 647 thousand traffic-accident reports collected over a period of four years by the New York city Police Department, in order to predict the locations of traffic accidents. They do so with a prediction accuracy of 78%. This paper borrows heavily from the ideas proposed in this research, but also develops a CNN based model architecture to make use of other structured data too.

4 FEATURE ENGINEERING AND ANALYSIS

The **Road Traffic Accidents** data set provides detailed road safety data about the circumstances of personal injury road accidents in Great Britain from 2018 to 2020, the types and classes of roads and junctions involved, the speed limit on the road of the accident location, climatic and weather conditions, the complete timestamp of the accident and the consequential casualties. The statistics relate only to personal injury accidents on public roads that are reported to the police, and then subsequently recorded. The **Census, 2011** data of Great Britain is the latest open population based open data set available and thus has been used to provide information about how densely crowded each region is. This has a correlation with the traffic accidents as it indicates the intensity of activity and chaos in a region. Since the event of an accident is not only dependent on external environmental factors, but also on the other vehicles present on the road during the event, incorporating relevant traffic data would improve the separability power of the data set to classify and differentiate accident prone feature vectors from the safe ones. Thus, the **AADF Traffic Data** is used which indicates the annual average daily flow (AADF) of traffic measured at various points on major and minor roads across the UK, according to the type of vehicle. The complete data set is an amalgamation of these 3 data sets, with a target variable that indicates the three classes of accidents, namely - **'Slight'**, **'Serious'** and **'Fatal'**.

In order to understand the various features and their interplay with the target classes, each feature and target pair is analysed separately.

4.1 Road based Features

These consist of Road Class, Road Type, Junction Details, Speed Limit and Urban or Rural Area. These help provide information about the finer details of the location of accidents that may act in conjunction to help identify patterns and alert the algorithm about potential risks of an accident. For example: The predictive algorithm may identify that A(M) motorway roads with a roundabout or a staggered junction with damp road surface conditions are more prone to severe accidents than their counterparts. **Road Class** is categorical feature consists of 5 categories (A, B, C. Motorway and Unclassified) that indicate the class of the road on which the accident took place. In order to comment on the risk of traveling on a particular road, consider the P(Accident Severity/Road Class) - Probability of Accident Severity given a Road Class, which highlights the investigative nature of the study by pointing towards the probability of the effect, given the cause. On computation of these conditional probabilities, it is found that Motorways are the most prone to fatalities (2.1%) as compared to its counterparts. **Road Type** indicates any one of the categories of roads - Single carriageway, Dual carriageway, Roundabout, One way street and Slip road. Upon similar treatment and computing the conditional probability of P(Accident Severity/Road Type), it is found out that Dual carriageways are more prone to fatalities than its counterparts (17%). On computing the class conditional probability for **Junction Details**, it is observed that roads having no junction within 20 meters have the most number of accidents and are the riskiest too, with respect to the probability of having a fatality. **Speed Limit** analysis suggests that 60MPH speed limit roads have highest probability of fatal accidents, even though 30MPH roads have the highest frequency of accidents. This can be explained by the safe assumption that the average vehicle speeds have a direct positive correlation with the speed limit of the road. But 70MPH roads have lesser probability of fatal accidents and a significant jump in probability of slight accidents, maybe due to the fact that 70MPH roads are particularly built and engineered in a way that makes driving at high speed a more comfortable and safer experience.

4.2 Time based Features

These consist of Date, Time and Day of the Week of the Accident events. In order to understand the seasonality and periodicity (if any) of the number of accidents with respect to the date and time, the time series of the accident numbers is decomposed into the trend, seasonality and the residuals as shown in Fig. 1. It can be seen from this monthly data series that there is a negative trend in the total number of accidents from 2018 to 2020 and a clear seasonal pattern can be detected. The cases rise sharply during the ending months of summer in July and August, then fall suddenly during September and shoot up to a maximum towards the end of the year in November and December. This repeated pattern (observed over the course of 2 years) has a direct correlation with the festive seasons. The Augmented Dickey-Fuller test of stationarity outputs the p-value as 0.0233 (<0.05) which means that the null-hypothesis of assuming that the data is stationary, is rejected.

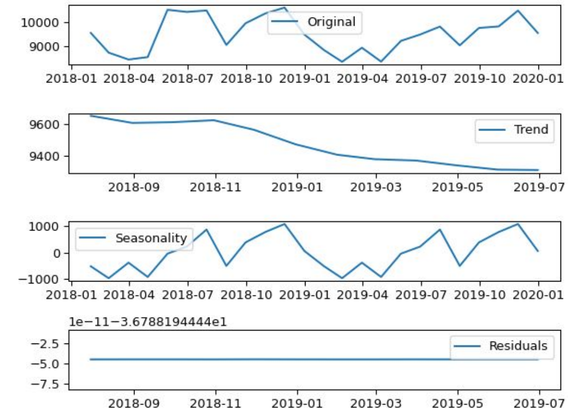

Figure 1: Decomposition of Accident Data Time Series (Accident counts vs Months)

Furthermore, by analysing the **Time** of accident events vs its counts, there is a clear trend to be noticed (see 2). The accident counts peak during the rush hour timings of 7AM - 9AM and 4PM - 7PM. Computing the class conditional probabilities of the feature Day of the Week results in Saturdays and Sundays having the highest probability of Fatalities (1.6% and 1.7% resp.).

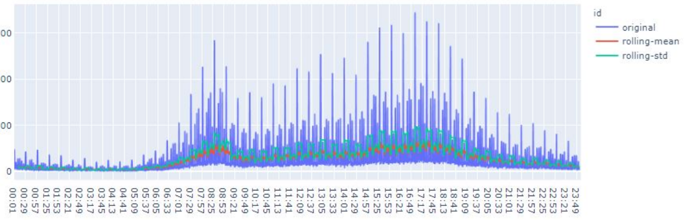

Figure 2: Accident counts vs Time of event

Thus, the timestamp based features have been transformed from categorical to cyclical features using their periodicity.

$$Time = \sin(\frac{Time \times 2\pi}{24 * 3600})$$

$$Day = \sin(\frac{Day \times 2\pi}{7})$$
$$Date = \sin(\frac{Day \times 2\pi}{365})$$

The time used in the equation is treated in seconds and the Day as the total number of days passed in the year, for maximum granularity in the datum of these features. Since a majority of the features in this data set are categorical variables, these continuous variables with increased granularity add to the local variance of the distribution of classes. Also, by transforming these time based features using the sinusoidal function, the data set not only avoids the curse of dimensionality (as Day and Date would have to be treated as categorical and thus One-Hot transformed), but also adds relational information to the features as now the representation of the day Sunday and the Dates in December are closer in magnitude to the day Monday and the Dates in January. This adds an extra layer of information.

4.3 Traffic Data and Population Density

For the most accurate data coherence, it would be apt to use hourly collected traffic data at each count point located in the road and highway networks. But since the hourly traffic data collected by the government of U.K. is only available for the peak traffic hours between 6AM - 6PM, it is inefficient to interpolate more traffic count values than actually present for further use. Thus, AADF data is used for the years 2018 and 2019. These contain the vehicle counts of two wheeled motor vehicles, cars, buses, taxis, LGVs and HGVs. All of these are counted at certain counting points throughout the country. Thus, to correctly merge these traffic counts with the accident data, the coordinates of the traffic count points and the accident events are used. Essentially, the nearest neighbour counting point of each accident location must be found and the data for the same must be merged with the accident data. But a brute force search on 240171x251352 accident locations x count points would be extremely time consuming. Thus, the KDTree algorithm using the sliding midpoint rule [15] is used to calculate the nearest neighbours across all the pairs of coordinates.

Population Density is calculated using the Census 2011 data that indicates the population and area (sq.hectares) of each LSOA (Lower Layer Super Output Area are a geographic hierarchy designed to improve the reporting of small area statistics in England and Wales). Finally, the traffic counts for each vehicle type and the population density have been log transformed since the distribution of each followed a power law graph. This makes each of these count based features follow a more Gaussian distribution.

4.4 Satellite Images

Using images offer multiple benefits

- By default, they encode the information available in structural format in a much more information rich manner. For example: an image of an accident site may depict the presence of a junction or a crossroad, the road surface conditions in some capacity, the pedestrian crossing information etc. This not only replaces the inefficient representation of information in the form of one-hot encoded vectors, but also adds onto it by acting as a proxy indicator for many more additional unexplored factors that may help predict the chances of an accident.
- These images can be combined with the remaining structural data of the traffic conditions. Time-stamp related attributes can't be used anymore since the 'safe' class images lack that information. But the traffic counts have a fair correlation with the time and the day of the week to incorporate the time-stamp information in a limited capacity

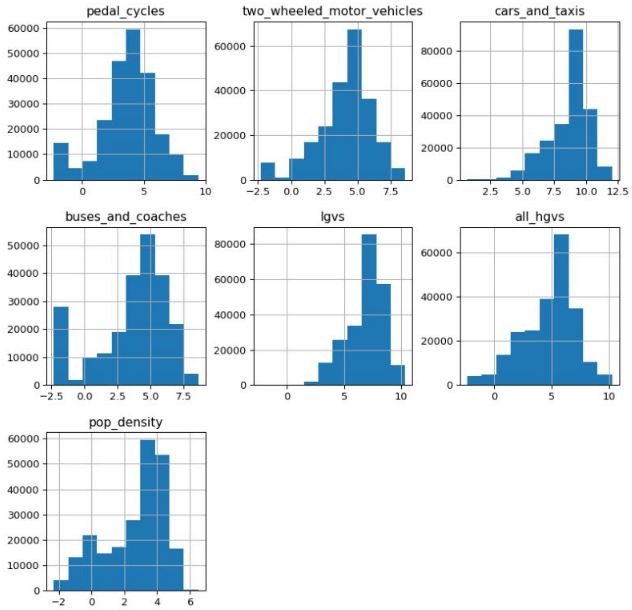

Figure 3: The distribution of all Traffic Counts and Population Density features after log transformation

Since accident data set only provides information about the locations and events when an accident occurs, there is a lack of data for the neutral class of 'No Accident'. After all, a user is only alerted when there is a significant risk of an accident. But for the classifier to differentiate between an accident datum and a 'safe' datum, it must be fed sufficient data points of the 'safe' class. To solve this problem, satellite imagery of the locations that don't appear in the accident data can be used.

Firstly, due to the monetary limit on HTTP requests to the Maps Static API, only the coordinates of London are used in order to carry on the predictive analysis. Thus, all the satellite images and the structured data used will be limited to the coordinate bounds of London with - Latitude ∈ [51.257, 51.719] and Longitude ∈ [-0.542, 0.291]. Now, the entire grid space of these (Lat, Long) bounds is divided into grid squares of 30mtr × 30mtr [18], as shown in 4. It is also important to round off the generated coordinates and the data set coordinates to the same decimal places in a methodical manner for better concurrence, since the 'Safe' class coordinates shall be calculated by computing a set subtract operation between the entire coordinate set of London binding 30 × 30 sq.mtr. spaces and the coordinate set of all the accidents that took place in London between 2018 and 2020. These coordinates collected for accident

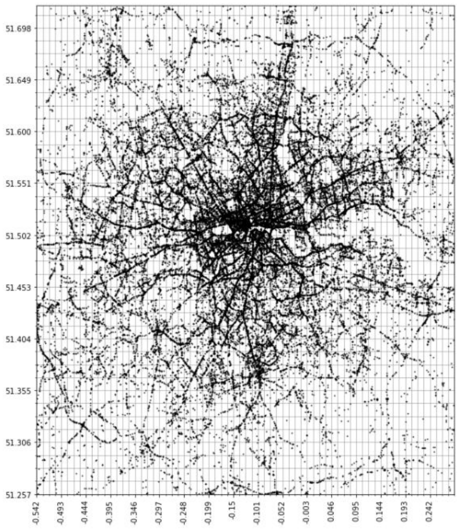

Figure 4: Grid Squares over the coordinate bounds of London with scatter points of Accident Locations

locations and safe locations are then used to send HTTP requests to the Maps Static API to download satellite images for the two said classes. A total of 6120 randomly sampled coordinates are used to download images for each class.

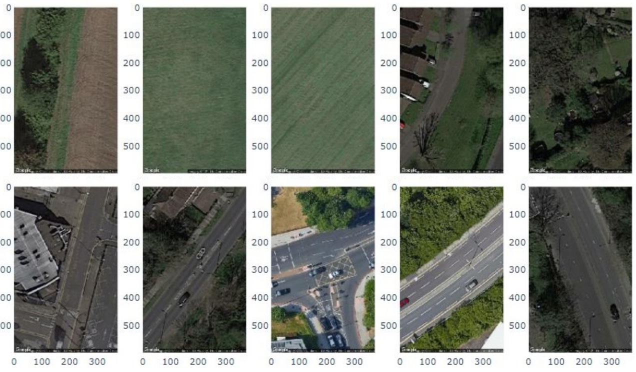

Figure 5: Satellite Images of Safe locations(top) and Accident Locations(bottom)

5 ARCHITECTURE

5.1 Light Gradient Boosting Framework for Structured Data

Using the data set prepared by the Road Accident, Road Traffic and Population density data sets, the learning task here is a binary classification task with the classes being 'Slight' and 'Severe' accident. The original data set contains three classes as mentioned in section 1, but due to the extremely small ratio of 'Serious' and 'Fatal' accident datums as compared to the 'Slight' accident datums, the former two are combined to form one class - 'Severe'. Even after doing so, the class ratio of Slight:Severe = 0.796:0.204. Thus, instead of down-sampling the majority class and losing valuable data, **SMOTE + ENN based Up-Sampling** is used [5]. The oversampling of minority classes may only work in the case when all the classes are not skewed and don't extend into the other classes. In the case that either of the majority or minority classes skew into the other ones, SMOTE would in fact lead to much poor performance of the classifier as it would add more noise into the previously well classified majority classes. To avoid this, an improved version of Tomek link removal [5] is used in conjunction with SMOTE - ENN. Since a majority of the features used in this data set are categorical and must be one-hot encoded, a decision tree based learning algorithm is chosen to handle all the 45 dimensions (after one-hot encoding). To this end, LightGBM [1] is used.

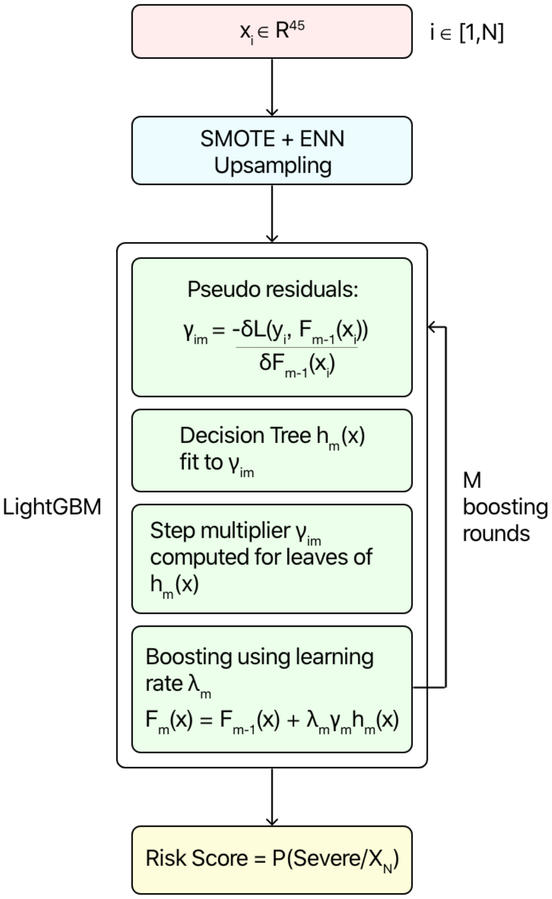
$x_i \in R^{45}$
$i \in [1,N]$
SMOTE + ENN Upsampling
Pseudo residuals:
$\gamma_{im} = \frac{-\delta L(y_i, F_{m-1}(x_i))}{\delta F_{m-1}(x_i)}$
Decision Tree $h_m(x)$ fit to γ_{im}
Step multiplier γ_{im} computed for leaves of $h_m(x)$
Boosting using learning rate λ_m
$F_m(x) = F_{m-1}(x) + \lambda_m \gamma_m h_m(x)$
LightGBM
M boosting rounds
Risk Score = P(Severe/X_N)

Figure 6: LightGBM predictor for structured data

5.2 Convolutional Neural Network for Image Data

Satellite Images downloaded for the 'Safe' and 'Accident' classes are of 600×375 pixel size, but are later transformed into 64×64 pixels due to faster processing. A CNN [13] based architecture is deployed to classify the images into the two classes. For better performance, instead of using custom made convolution blocks, VGG19 [19] is

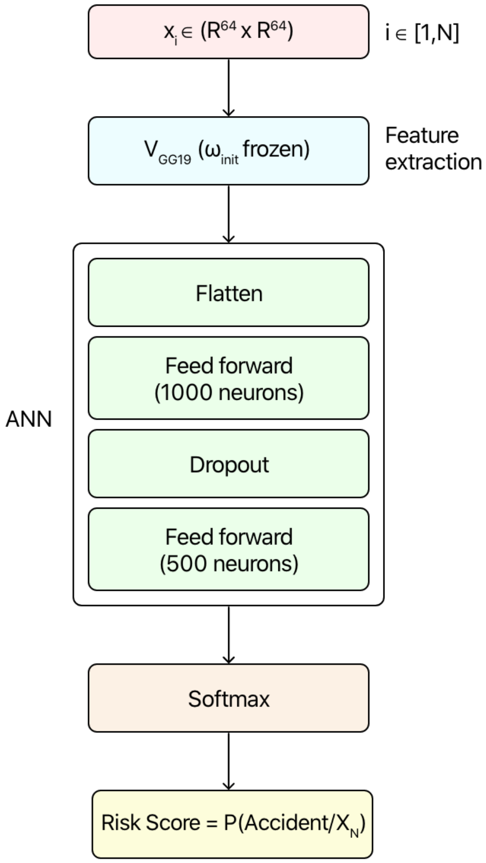

Figure 7: VGG19 + ANN predictor for Satellite Image data set

used with its weights initialised to the the pre-trained ones on the ImageNet [7] data set. This helps in *Transfer Learning* the features and activations from the images and then passing it onto a fully connected neural network for binary classification.

5.3 Sequential-Ensemble Learning for Heterogeneous Data (proposed)

The idea behind deploying an ensemble learning model is to make use of both, the satellite images to include the external environmental and road condition factors and the AADF Road Traffic data for vehicle traffic information based on geo-location to improve the predictive capability of the algorithm. Here, the probability scores being generated by the VGG19 + ANN block for the satellite images are used as subsequent features to be combined with the AADF traffic counts at the respective locations. This data merging is done using the same methodology of KDTrees based nearest neighbour computation, as described in section 3. This structured data set is then used as an input for the LightGBM decision tree model to predict the final probability scores of 'Safe' and 'Accident' class.

The **Risk-Score** is generated using the probability output for the 'Severe'/'Accident' class by the Model. The user is alerted when this risk score exceeds 0.5 for a continuous duration of *t* seconds, indicating that the data being collected indicates a greater probability of a 'Severe' accident than a 'Slight' accident or 'Safe' location.

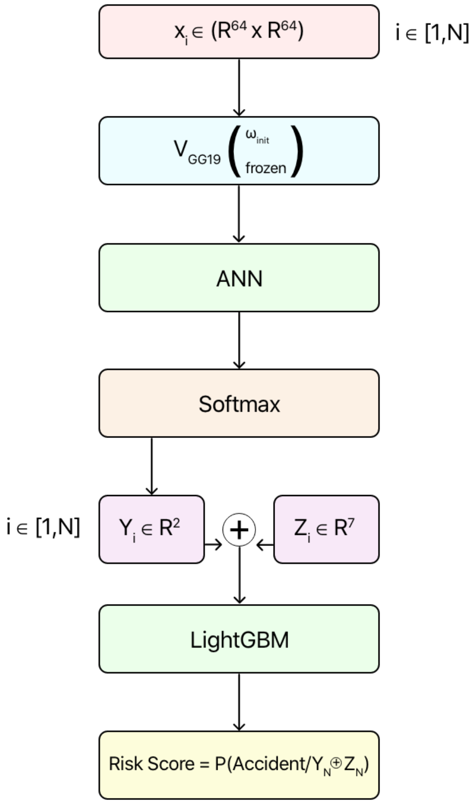

Figure 8: Ensemble Learning for Image + AADF + Population density data set

6 RESULTS AND CONCLUSIONS

The classification architectures are built and deployed on their respective data sets to obtain the results as shown in Table 1. LightGBM on its own does not produce satisfactory results for the classification task due to a huge class imbalance, even after extensive hyper parameter tuning. SMOTE + ENN over sampling helps the LightGBM model perform better, but not significantly. This indicates that the method of oversampling by adding additional scaled

Architecture	Train-Acc.	Val-Acc.	Test-Acc.
LightGBM	55%	54.5%	50.3%
LightGBM-SMOTE+EMM	62%	60%	58%
VGG19 + ANN	84.92%	78.86%	76.67%
(VGG19+ANN)-(LightGBM)	-	-	-

Table 1: Accuracy of each Architecture

datum-vectors along the mid-points of pairs of vectors belonging to the minority class only works when the inherent distribution of the class is regularly distributed in a well defined region, already represented/occupied by the limited available datums. Additionally, even after oversampling the minority classes, the ratio of class-1 samples against class-0 samples only improves by a slight margin. This is because ENN removes noisy samples that reduce class separability from both the classes irrespective of their ratio against the others. This indicates that oversampling caused a huge number of samples to be generated in the space already occupied by the distribution of the majority class (class-0). Thus, almost zero-to-none information was added to the minority class for the classifier to improve on and the results remained the same. This is indicative of the fact that the information/data available is not fit for a generalizable classification task, unless coupled with more features or additional datums. In this light, it is clear that the Satellite Image based CNN model performs much better than the LightGBM model for structured data. This is due to the extra dimensions of information added on by the images. The Sequential Ensemble Learning model results haven't yet been computed and have been kept as placeholders, in the hope that it uses the independent class separability power of both the data sets and improves the classification accuracy even further.

7 FUTURE WORK

Possible future work could focus on the improvement of these models using better, more sophisticated transfer learning and feature extraction models such as ResNet50 [9], and training these on more extensive data sets. An attention based architecture could also be added to these CNN blocks for an improved detection of the useful features and activations from the images, as used by [8], but using a mixed input data model by adding on structured data. Lastly, by generating these risk scores for locations on road networks based on a myriad number of features, this problem could be translated into a Multi-Agent Reinforcement Learning problem where the goal is to design efficient road traffic management systems that minimize the cumulative probability/risk of accident for all the users on a given road segment.

REFERENCES

[1] Welcome to lightgbm's documentation!

[2] Global status report on road safety 2015, Dec 2018.

[3] Abdel-Aty, M. A., and Radwan, A. Modeling traffic accident occurrence and involvement. *Accident Analysis Prevention 32*, 5 (2000), 633 – 642.

[4] Cadamuro, G., Muhebwa, A., and Taneja, J. Assigning a grade: Accurate measurement of road quality using satellite imagery. *ArXiv abs/1812.01699* (2018).

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res. 16*, 1 (June 2002), 321–357.

[6] Das, P., and Chand, S. Extracting road maps from high-resolution satellite imagery using refined dse-linknet. *Connection Science* (2020), 1–18.

[7] Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255.

[8] Gupta, S., Srivatsav, D., Subramanyam, A. V., and Kumaraguru, P. Attentional road safety networks. *2019 IEEE International Conference on Image Processing (ICIP)* (2019), 1600–1604.

[9] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

[10] He, S., Bastani, F., Jagwani, S., Park, E., Abbar, S., Alizadeh, M., Balakrishnan, H., Chawla, S., Madden, S., and Sadeghi, M. Roadtagger: Robust road attribute inference with graph neural networks. *ArXiv abs/1912.12408* (2020).

[11] Kamalu, J., and Choi, B. Road mapping in low data environments with openstreetmap. *ArXiv abs/2006.07993* (2020).

[12] Kornfeld, N., Lücken, L., Leich, A., Wagner, P., Saul, H., and Hoffmann, R. Crash rate estimation by aerial image analysis.

[13] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation 1*, 4 (1989), 541–551.

[14] Lenk, A., Cersosimo, M., and Raoof, N. A novel approach for predicting and understanding road danger in the developing world: Deep video-classification of roads in nairobi, kenya.

[15] Maneewongvatana, S., and Mount, D. M. Analysis of approximate nearest neighbor searching with clustered point sets. *CoRR cs.CG/9901013* (1999).

[16] More, A. Survey of resampling techniques for improving classification performance in unbalanced datasets, 2016.

[17] Nachmany, Y., and Alemohammad, H. Detecting roads from satellite imagery in the developing world. In *CVPR Workshops* (2019).

[18] Najjar, A., Kaneko, S., and Miyanaga, Y. Combining satellite imagery and open data to map road safety. In *AAAI* (2017).

[19] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2015.

[20] Transport, D. f., Oct 2020.