Predictive Analysis of Accidents Based on US Accident Data

Mayura Manawadu
School of Electronics Engineering
Kyungpook National University
Daegu, South Korea
mayuramanawadu@knu.ac.kr

Udaya Wijenayake
Department of Computer Engineering
University of Sri Jayewardenepura
Nugegoda, Sri Lanka
udayaw@sjp.ac.lk

***Abstract*— Traffic accidents have emerged as a serious global concern, resulting in daily casualties and prompting authorities to prioritize accident prevention measures. This study presents an accident prediction system that alerts drivers of potential accidents by analyzing multiple attributes which are the potential courses of accidents. While previous studies have predominantly focused on analyzing geographical factors, predicting accident frequencies, and assessing accident risks, this study aims to develop two systems—an advanced route recommendation system and a real-time accident prediction system—to enhance road safety. Moreover, existing systems often predict post hoc accident occurrences or have limited geographical coverage. The advanced route recommendation system is designed to assist users in planning their journeys by providing them with the safest routes in advance. Through a website interface, users can log in and receive personalized recommendations on the accident prone areas on their path, based on factors such as historical accident data, road conditions, traffic patterns and weather conditions. This system aims to empower individuals to make informed decisions and reduce the likelihood of accidents during their trips in advance. The real-time accident prediction system aims to provide drivers with up-to-date information on potential accidents along their routes. By utilizing GPS coordinates and retrieving live data, including weather conditions and real-time accident reports, the system predicts accident-prone areas in real-time. Drivers receive these predictions through a mobile application as an audio message, enabling them to make timely adjustments to their routes and avoid hazardous situations. Additionally, static predictions are displayed on a website, featuring markers indicating accident-prone areas. The research utilizes an extensive dataset of "US Accident Dataset" spanning across all 49 states of the USA. Results demonstrate that the Random Forest classifier achieves an impressive 91.5% accuracy in predicting accident severity, surpassing previous studies. Furthermore, this paper conducts Exploratory Data Analysis, unveiling intriguing patterns in the dataset regarding accident occurrences.**

Keywords*— accident prediction, data science, real-time systems, Random Forest Classifier, Exploratory Data Analysis.

I. INTRODUCTION

Road crashes continue to pose a significant global challenge, with alarming statistics indicating that someone loses their life every 24 seconds due to road accidents, resulting in an estimated 1.35 million fatalities annually [1]. Additionally, the World Health Organization reports that 20-50 million individuals sustain injuries from road crashes each year [2]. The United States, in particular, experiences a significant toll, with over 38,000 annual deaths and 4.4 million severe injuries caused by road accidents. Tragically, road crashes are the leading cause of death among U.S. citizens aged between 1-54 [3]. The financial impact of these accidents is also substantial, with medical costs amounting to over 380 million dollars annually. Among high-income countries, the US faces the highest number of road crash fatalities [3].

In response to these grave statistics, authorities have implemented various precautionary measures to mitigate road accidents, including the display of warning messages through signage, installation of traffic signs, and the use of road bumps, among others. However, most of these actions have primarily focused on static factors such as location data. Dynamic factors, such as weather conditions (e.g., rainfall, precipitation, temperature, wind speed, humidity, and pressure) and time-related variables (e.g., day time, night time), can also act as catalysts for traffic accidents. Integrating the power of Data Science into accident analysis and prediction can be instrumental in effectively addressing these factors. Therefore, it is crucial to consider all relevant variables in the predictive analysis of accident causation. Road accidents seldom occur spontaneously; they often exhibit discernible patterns that can be predicted and prevented. As such, accidents represent events that can be examined, analyzed, and mitigated.

While several existing accident prediction systems have been developed, this study aims to present a unique and robust accident prediction system that sets itself apart from previous approaches. Unlike many existing systems that predominantly consider static factors, our system comprehensively incorporates a wide range of potential stimuli for accidents, including weather data, location data, time-series data, and more. By leveraging the rich dataset provided by the US Accidents Dataset [4] , our approach captures a holistic view of the factors influencing accidents.

Accident Prediction has become a promising branch among the AI research community. The purpose of this study is to present a robust accident prediction system that considers a wide range of possible stimuli for accidents. Thus the presented work in this paper is conducted based on the US Accidents Dataset which is an enriched dataset comprised of weather data, location data, time-series data, etc. From this paper, we present a real-time accident prediction system using different paradigms of machine learning. The trained model is deployed in Google Cloud and connected the front end with a mobile device to make real-time predictions. The mobile device will communicate with the server to receive prediction results by sending current GPS coordinates to the server. Simultaneously a website is also designed to make static predictions within 48 hours ahead of the present time which

acts as a route recommendation system. This paper also presents possible patterns and factors leading to the accidents which are extracted by instituting different data mining techniques.

While there have been notable contributions in the field of accident prediction, the presented work offers unique contributions and differentiates itself from existing studies in several ways. First, unlike previous studies that primarily focused on specific states, cities, or small geographic areas, this research encompasses the entire United States of America, providing a broader perspective on accident prediction. Moreover, our system incorporates dynamic environmental stimuli, including weather conditions and time data, which have been shown to significantly impact accident occurrences. By leveraging a comprehensive dataset and considering these influential factors, we aim to enhance the accuracy and effectiveness of accident prediction, ultimately contributing to proactive accident prevention strategies.

II. Related Work

The field of predictive and descriptive analysis of accidents using past data has garnered significant research attention, driven by the emergence of diverse data science paradigms. Numerous studies have been conducted in this domain, making it a focal point within the research community. This section reviews the state-of-the-art works related to the approach presented in this paper.

Sobhan Moosavi et al. [5] proposed a deep neural network model, named DAP, for real-time accident risk prediction based on the US Accidents Dataset, which they also curated. Their work introduced a deep neural network architecture that incorporated recurrent, embedded, and fully connected components. All attributes present in the dataset were considered. The maximum F1 score achieved for the label class was 0.65. Although their model demonstrated promising results, higher accuracy could have been attained through the implementation of more advanced preprocessing and data augmentation techniques.

Honglei Ren et al. [6] propose a Deep Learning approach, based on Long Short Term Memory (LSTM) for predicting traffic accident risks. Their model achieved improved accuracy, with a Root Mean Squared Error of 0.034. Through pattern analysis, they revealed that traffic accidents were not distributed uniformly in space and time. However, the dataset utilized in their study was limited in scope, primarily focusing on a specific set of constraints and emphasizing time series analysis.

Lu Wenqi et al. [7] proposed a traffic accident prediction model called TAP-CNN, which utilized a Convolutional Neural Network (CNN) architecture with two hidden layers and a ReLU activation function. The TAP-CNN model achieved an accuracy of 78.5% in its predictions. The authors employed United States I-15 highway 160 mile-166 mile road accident data to evaluate their model. However, it is worth noting that their dataset was restricted to a small geographic area within the United States.

Senk et al. [8] have researched the use of accident prediction models in identifying hazardous road locations. Their study focused on utilizing historical accident data and analyzing spatial patterns to identify high-risk areas. However, it is important to highlight that their work did not consider dynamic factors such as weather conditions, time data, and other environmental stimuli, which can significantly influence accident occurrences. But their work does not take dynamic factors such as weather conditions, time data, etc. into consideration.

Tessa K. Anderson proposes an approach to profile hotspots of accidents using K-means Clustering [9]. By leveraging accident data collected by the Metropolitan Police of the UK, the study employed a kernel-based density tool to visualize accident events based on density. This approach provided valuable insights into accident patterns and hotspots within a specific region. While designing an accident prediction system, a crucial consideration is whether to develop a system tailored to specific regions, cities, or states or to design a more generalized system applicable across diverse areas. A region-specific approach may offer advantages in terms of fine-tuned predictions, catering to local conditions and user preferences. On the other hand, a generalized system, which is the focus of our work, brings the potential for widespread impact, collaborative efforts, and the accumulation of insights from a broader range of data sources.

In contrast to the aforementioned studies, the presented work in this literature aims to enhance accident prediction by employing various data management techniques on the comprehensive US Accidents Dataset. The authors focus on developing a real-time accident prediction system that incorporates dynamic environmental stimuli, leading to higher accuracy. Importantly, the system's scope is not limited to a specific state or city but encompasses the entire United States of America, providing a broader perspective on accident prediction. By leveraging this comprehensive dataset and considering dynamic factors, the proposed system aims to advance the field of accident prediction and contribute to more effective accident prevention strategies.

III. Methodology

A. Dataset

The dataset used in this research is based on the US Accident dataset, which is continuously updated with accident records obtained through the MapQuest and Bing APIs. For this study, a snapshot of the dataset from February 2016 to June 2020, containing 4 million accident records, was taken. The dataset encompasses a wide range of attributes, including time data, location data, weather data, and point of interest (POI) data, as outlined in Table I.

B. Exploratory Data Analysis

Having 4 million records of accidents in the dataset, it was possible to extract interesting patterns related to the accidents using different data mining techniques.

1) Statewise accident distribution: Fig. 1 demonstrates that California records the highest number of accidents, followed by Texas and Florida.

2) City-wise accident distribution: Fig. 2 highlights that Houston, Los Angeles, Charlotte, Austin, and Dallas are the top 5 cities with the highest number of accidents.

3) Distribution of accidents based on severity: Fig. 3 presents a scatter plot of accidents based on latitude and longitude, with darker points indicating more severe accidents. The plot shows that the eastern part of the USA is more prone to severe accidents, while the western part predominantly experiences medium-level accidents.

TABLE I ATTRIBUTES OF THE DATASET

Attribute	Description
ID	Unique identifier of the accident record
Severity	Severity level of the accident from 1 to 4
Start_Time	Start time of the accident in local time zone
End_Time	End time of the accident in local time zone
Start_Lat	Start latitude of the accident
Start_Lng	Start longitude of the accident
End_Lat	End latitude of the accident
End_Lon	End longitude of the accident
Distance	Length of traffic affected by the accident
Number	Shows the street number in the address field
Street	Street name in address field
Side	Relative side of the stress (Right/Left)
City	City in the address field
Country	Country
State	Shows the state in address field.
Zipcode	Shows the zip code in address field.
Country	Shows the country in address field.
Timezone	Timezone of location (eastern, central, etc.).
Airport_Code	Closest airport near to the location
Weather_Timestamp	Time-stamp of weather observation record
Temperature(F)	Temperature (in Fahrenheit).
Wind_Chill(F)	Wind chill (in Fahrenheit).
Humidity(%)	Humidity (in percentage).
Pressure(in)	Air pressure (in inches).
Visibility(mi)	Shows visibility (in miles).
Wind_Direction	Direction of the wind
Wind_Speed(mph)	Wind speed (in miles per hour).
Precipitation(in)	Precipitation amount in inches
Weather_Condition	Weather Condition (rain, snow, , fog, etc.)
Amenity	Indicates the presence of Amenity nearby
Bump	Presence of speed bump or hump
Crossing	Presence of crossing in a nearby location.
Give_Way	Presence of give_way in a nearby location.
Junction	Presence of junction in a nearby location.
No_Exit	Presence of no_exit in a nearby location.
Railway	Presence of railway in a nearby location.
Roundabout	Presence of roundabout in a nearby location.
Station	Presence of station in a nearby location.
Stop	Presence of stop sign in a nearby location.
Traffic_Calming	Presence of traffic_calming sign neaby
Traffic_Signal	Presence of traffic signal in a nearby.
Turning_Loop	Presence of turning loop in a nearby location.
Sunrise_Sunset	Period of day based on sunrise/sunset.
Civil_twilight	Period of day based on civil twilight.
Nautical_twilight	Period of day based on nautical twilight.
Astronomical_twilight	Period of day based on astronomical twilight.

Our decision to design a generalized accident prediction system covering stems from our goal to ensure accessibility and applicability across various regions, cities, and states. By creating a system that transcends geographical boundaries, we aim to contribute to broad-scale accident prevention measures and facilitate collaboration among authorities.

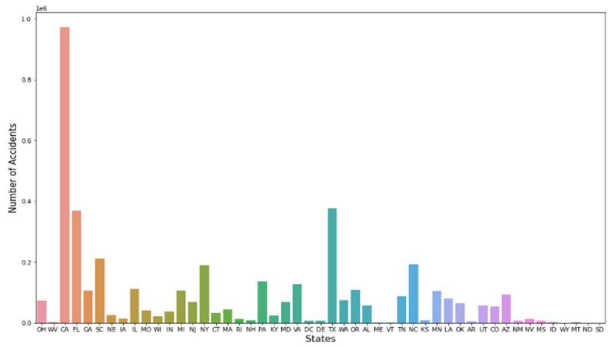

Fig. 1. Count of Accidents by State.

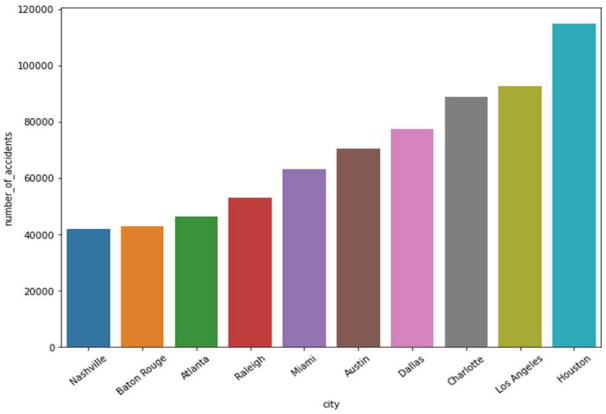

Fig. 2. Count of Accidents by City.

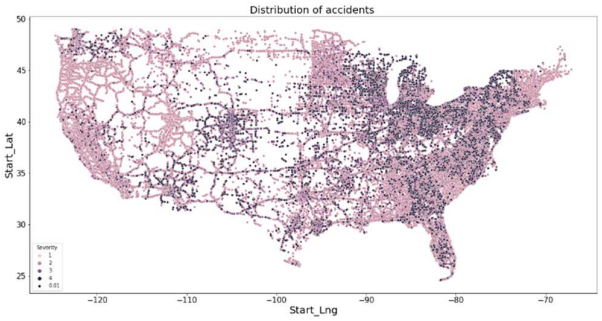

Fig. 3. Scatter plot of Accident Distribution in the USA.

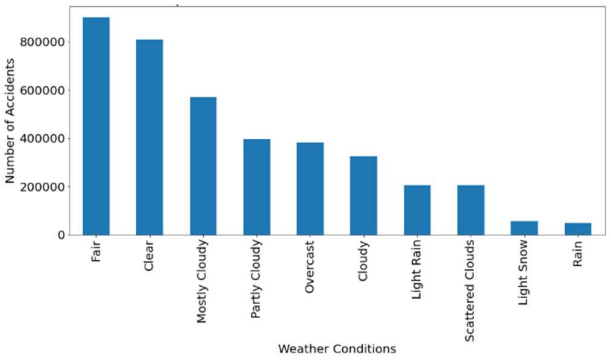

Fig. 4. Number of Accidents based on Weather.

4) Accidents based on Weather: Fig. 4 groups the number of accidents based on weather conditions, indicating that most accidents occur under fair weather conditions. However, it is important to note that the graph alone does not provide a conclusive impact assessment as the occurrence of rain, snow, and other extreme weather conditions is relatively rare compared to fair weather conditions.

5) Analysis of Time Series Data: Exploratory data analysis based on time series reveals interesting insights. Fig. 5 demonstrates an increasing trend in the number of accidents over the years, potentially attributed to the rising vehicle population. Fig. 6 displays monthly accident rates, showing an elevation at the end of the year, likely due to the onset of winter and snowfall.

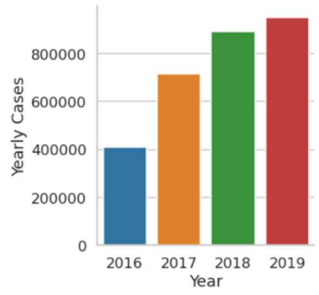

Fig. 5. Yearly Accident Distribution.

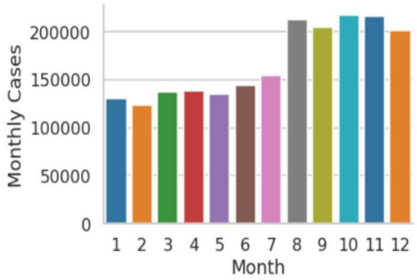

Fig. 6. Monthly Accident Distribution.

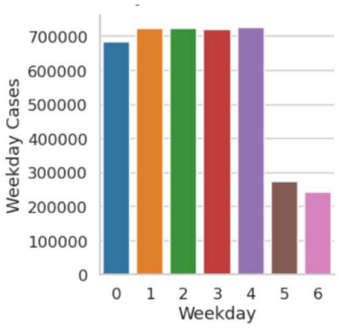

Fig. 7. Daliy Accident Distribution over week.

Fig. 7 reveals that most accidents occur on weekdays, while weekends report relatively fewer accidents, potentially due to heavy traffic during weekdays. Fig. 8 indicates that the highest hourly accident rates occur between 06:00-09:00 and 16:00-18:00, corresponding to morning and afternoon traffic peak hours when people commute to offices and schools, resulting in higher vehicle density.

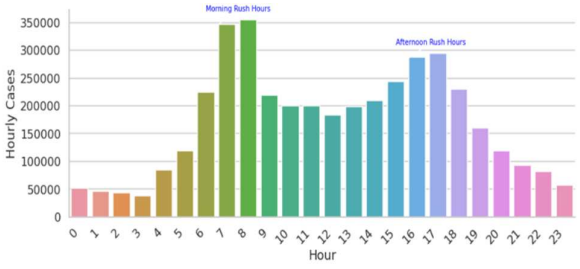

Fig. 8. Hourly Accident Distribution.

C. Data Cleaning and Preprocessing

The dataset underwent data cleaning and preprocessing steps. Highly correlated features were dropped using a correlation matrix generated with Matplotlib. Additionally, irrelevant features such as ID, End Time, End Latitudes, End Longitudes, and accident duration were removed as they provided no useful information for future predictions. Missing values were handled using different approaches: dropping features with more than 50% missing values, dropping null values when the percentage was less than 1%, and employing imputation techniques such as replacing missing categorical data with the mode and missing continuous data with the median. Time series features were normalized and augmented into Year, Month, Date, Hour, and Minute attributes. Weather attributes with multiple categories, such as wind direction, were transformed into broader categories. Weather conditions were grouped into six main groups: Cloudy, Clear, Rain, Heavy Rain, Snow, and Fog, allowing for more generalized values for each attribute to improve predictions. The Time Series features from the dataset were normalized and augmented into Year, Month, Date, Hour, Minute attributes. The attributes like wind direction which had variables of 8 categories were transformed into four main directions. As an example, WSW (West-Southwest) and WNW(West-Northwest) were transformed into West. Similarly, the data set consists of a wide variety of weather conditions. Those weather conditions were transformed into 6 main groups as Cloudy, Clear, Rain, Heavy Rain, Snow, and Fog. Through the transformations, a more generalized version of the values for each attribute was obtained for predictions.

D. Feature Engineering

To select and extract features from the dataset, several feature engineering techniques were employed. The location names included within the addresses were unique over the entire dataset. Therefore, those were labeled using their frequency encoding and log transformation as in (1).

$$x = log(x + 1) \quad (1)$$

The categorical data were subjected to One Hot Encoding so that they can be used to feed into the learning algorithms. Most of the learning algorithms cannot work directly with categorical data. The categories need to be converted into numbers.

E. Handling Class Imbalances

The US The US Accidents dataset exhibits imbalances among the class labels, with severity levels 1 and 2 dominating the recorded accidents (Fig. 9). To address this issue and avoid overfitting, two approaches were employed. The first approach involved upsampling and downsampling techniques to balance the number of data points in each class. The second approach utilized the Synthetic Minority Oversampling Technique (SMOTE), which generates synthetic data for the minority class using the k-nearest neighbor algorithm. Both approaches aimed to mitigate the imbalances and improve model development.

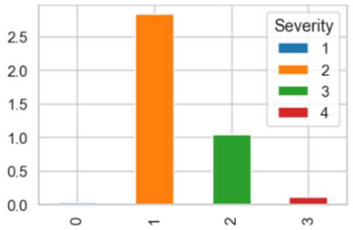

Fig. 9. Distribution of data points for each class.

F. *System Overview*

After applying the aforementioned approaches, the dataset was used to train different machine learning algorithms. The overall system flow is depicted in Fig. 10. Once a model with high accuracy was trained, it was deployed on a Google Compute Engine to establish a server accessible from anywhere in the USA. This deployment reduced the load on the frontend device by distributing the model weight to the server. Two front ends were developed: a website for static predictions up to 48 hours ahead and a mobile application for real-time accident detection. The front ends collect geographical coordinates and send them to the backend server for predictions. The mobile application reads latitude and longitude coordinates in real time, while the website collects start and destination locations and communicates with the backend server, which then uses the Google Routing API service to gather GPS coordinates along the route. The backend server further obtains corresponding weather coordinates and related data using the OpenWeather Maps API and Google Maps API using the received geographical cordinates. After gathering all the required data, predictions are made and displayed on the website using markers on Google Maps (Fig. 11). In the mobile application, predictions are provided through an audio message in addition to textual outputs, facilitating drivers' convenience and safety (Fig. 12).

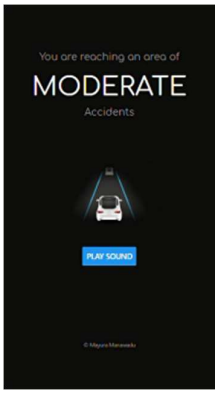

Fig. 12. Mobile Application Interface for Realtime predictions.

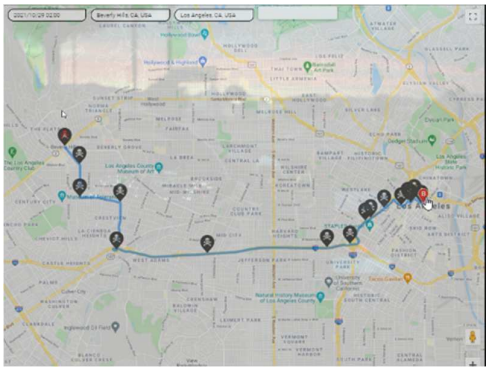

Fig. 11. Website designed for Static Predictions.

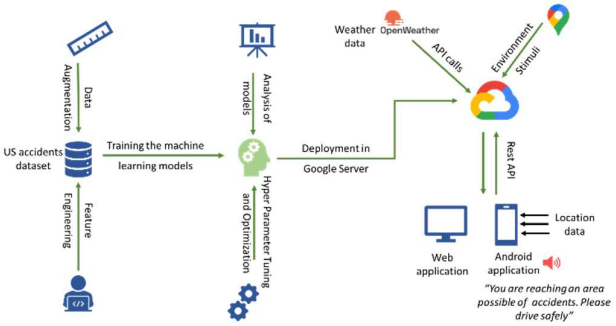

Fig. 10. Solution Architecture for the Accident Prediction System.

IV. Experiments and Results

The experimental results indicate that addressing the issue of class imbalance is crucial for improving the accuracy of the accident prediction system. Without balancing the dataset, the accuracy achieved was only 77%, which is insufficient for a critical system like accident predictions. Downsampling the dataset to an equal number of data points per class alone did not yield satisfactory results, likely due to the sparsity of the data points and the inability to cover the entire range of accidents in the USA.

To overcome these limitations, the dataset was downsampled to 27,000 data points for each class, which improved the accuracy. Additionally, the Synthetic Minority Oversampling Technique (SMOTE) was employed, resulting in a significant increase in accuracy to 90.8%. The Random Forest Classifier demonstrated the best performance among the tested classifiers, likely due to its ability to handle high-dimensional data and its use of subsets of features in model building.

Further optimizations were achieved by employing hyperparameter tuning and optimization techniques for the Random Forest Classifier. By setting appropriate values for parameters such as maximum depth, maximum features, and number of estimators, the accuracy was further increased to 91.5%.

It is important to note that the results obtained in this study are specific to the dataset and the machine learning algorithms used. The performance of the accident prediction system may vary when applied to different datasets or when using alternative machine learning algorithms. Additionally, the accuracy achieved in this study does not guarantee the absolute accuracy of the system in real-world scenarios. The accuracy reported should be interpreted as a measure of the model's performance within the experimental setup.

Despite these limitations, the results of this study demonstrate the effectiveness of addressing class imbalance and employing appropriate machine learning techniques for improving the accuracy of an accident prediction system. The achieved accuracy of 91.5% suggests the potential of the developed system to accurately predict the severity of accidents based on various attributes. Furthermore, it's noteworthy that the system's scope, whether generalized or personalized, can influence the dynamics of class imbalances. While our system targets a wide audience, the challenge of handling class imbalances may vary when applied to localized implementations. Future adaptations of the system could consider adjusting these techniques to cater to specific regions and thereby strike a balance between personalized predictions and addressing imbalances.

TABLE II ACCURACIES OF TRAINED MODELS

	Logistic Regression	Decision Tree	Random Forest	Multi-Layer Perceptron
Without Balancing	62	68	77	70.4
Down sampling (7000 Datapoints)	25	54	63.8	35.8
Down sampling (27000 Datapoints)	70.4	78.8	80.1	74.5
SMOTE	30	90.4	90.8	85.8

When comparing the performance of each classifier, the Random Forest Classifier yields the best accuracy in each approach. This could be due to Random Forest is performing well with high dimensional data since it involves subsets of data. It is faster than decision trees as it involves only a subset of features in a model. Therefore, Random Forest was selected for further optimizations.

V. CONCLUSION

In conclusion, this paper introduces a robust accident prediction system that achieves a high accuracy rate of 91.5% through the strategic utilization of diverse data management techniques and advanced machine learning algorithms. The system utilizes on the US Accident dataset and employs APIs to facilitate real-time predictions by capturing relevant attributes crucial for accurate forecasting. These predictions are efficiently relayed to drivers via audio output messages, ensuring prompt alerts concerning potential accidents. Complementing this system is a web application that empowers users to premeditate routes and assess potential accident-prone regions within the forthcoming 48 hours. Such insights can significantly aid authorities in the proactive implementation of safety measures within high-risk locales. Moreover, the exploratory data analysis methods employed have revealed intriguing patterns, enriching the understanding of accident occurrences.

Future enhancements could further amplify prediction accuracy by incorporating supplementary features that extend beyond the confines of the current dataset. Future directions might explore mechanisms to seamlessly integrate user preferences, including travel time and safety considerations, into our predictions. By striking a balance between generality and personalization, we can enhance the system's ability to empower users to make informed decisions tailored to their priorities.

On a broader scale, our research establishes a robust foundation for a generalized accident prediction system that possesses the potential to contribute significantly to accident prevention endeavors. While our focus has predominantly been on this generalized approach, the evolving landscape of machine learning and user-centric technologies beckons the exploration of more personalized systems. Potential directions may encompass seamlessly embedding user preferences, encompassing factors such as travel time and safety considerations, into our predictions. By navigating the fine line between generality and personalization, our system's efficacy in empowering users to make well-informed decisions tailored to their priorities can be further accentuated.

Furthermore, as a forward-looking trajectory, we propose the investigation of evolutionary algorithms, particularly Evolving Fuzzy Logics, to develop a prediction model adept at adapting to dynamically evolving environmental conditions. This advancement holds the promise of elevating the system's adeptness in addressing dynamic circumstances, thereby contributing to the refinement of accident prediction accuracy.

ACKNOWLEDGEMENTS

This research was supported by the Science and Technology Human Resource Development Project, Ministry of Education, Sri Lanka, funded by the Asian Development Bank (Grant No. STHRD/CRG/R1/SJ/06).

REFERENCES

[1] "Global road safety statistics," *Brake*. https://www.brake.org.uk/get-involved/take-action/mybrake/knowledge-centre/global-road-safety (accessed Nov. 07, 2021).

[2] "Road traffic injuries." https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries (accessed Nov. 07, 2021).

[3] "Road Safety Facts," *Association for Safe International Road Travel*. https://www.asirt.org/safe-travel/road-safety-facts/ (accessed Nov. 07, 2021).

[4] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath, "A Countrywide Traffic Accident Dataset," *arXiv:1906.05409 [cs]*, Jun. 2019, Accessed: Nov. 07, 2021. [Online]. Available: http://arxiv.org/abs/1906.05409

[5] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights," *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 33–42, Nov. 2019, doi: 10.1145/3347146.3359078.

[6] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction," *arXiv:1710.09543 [cs]*, Apr. 2018, Accessed: Nov. 07, 2021. [Online]. Available: http://arxiv.org/abs/1710.09543

[7] L. Wenqi, L. Dongyu, and Y. Menghua, "A model of traffic accident prediction based on convolutional neural network," in *2017 2nd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, Singapore, Singapore, Sep. 2017, pp. 198–202. doi: 10.1109/ICITE.2017.8056908.

[8] P. Šenk, J. Ambros, P. Pokorný, and R. Striegler, "Use of Accident Prediction Models in Identifying Hazardous Road Locations," *ToTS*, vol. 5, no. 4, pp. 223–232, Dec. 2012, doi: 10.2478/v10158-012-0025-0.

[9] T. K. Anderson, "Kernel density estimation and K-means clustering to profile road accident hotspots," *Accident Analysis & Prevention*, vol. 41, no. 3, pp. 359–364, May 2009, doi: 10.1016/j.aap.2008.12.014.