Contrasting Big Data Techniques in Exploring New Zealand Road Crash Data

Baosen (Edison) Hu
25290345
Auckland University of Technology
Auckland, New Zealand
jgc4857@autuni.ac.nz

Stephen J. Thorpe
9301663
Dept of Computer and Information Sciences, Auckland University of Technology
Auckland, New Zealand
stetho09@aut.ac.nz

***Abstract*—The present study explored the application of Big Data techniques, specifically Hadoop and MapReduce, to improve the analysis of the impact of weather and speed on motor vehicle crashes in New Zealand. Motor vehicle crashes result in high social and economic costs globally and in New Zealand. Therefore, accurate analysis of crash events is critical for evidence-based prevention and policy. In the present study, contemporary Big Data approaches were applied to address the limitations inherent in the traditional methods of crash analysis. We used Hadoop's distributed storage and MapReduce's processing capabilities on the New Zealand Transport Agency's Crash Analysis System (CAS) dataset to identify and visualize environmental and spatial trends to a higher degree of understanding. The project involved Elasticsearch and Kibana to make sense of unstructured data in geographic views, while Hue, Hive, and Power BI represented structured data with charts and dashboards. Results show that non-injury crashes, followed by minor crashes, are the most frequent, with over half happening at speed limits between 40–60 km/h. Geographically, Auckland represents crashes five times greater than in the other locations. Strong and extreme weather conditions appear to be a factor in the majority of reported fatal road accidents.**

Keywords—Big Data, Hadoop, MapReduce, Road Crash Data, New Zealand, Crash Analysis System, Weather, Speed Limit

I. Introduction

The estimated social cost for every fatal vehicle crash in New Zealand has been approximated at NZD4.9 million [1]. In the last ten years the average annual road toll has seen some 339 deaths each year, according to the Ministry of Transport [2]. In 2019, the fatality rate from road crashes was 7.1 per 100000 persons in New Zealand, compared to 5.1 in the European Union [3]. Fig. 1. below presents the total number of crashes by year between 2017 and 2024, which explains the importance of data-driven analysis for safety purposes. On top of this, the cost of injuries and other damages sees the country impacted by an estimated NZD14.9 billion in financial impact from motor vehicle accidents each year.

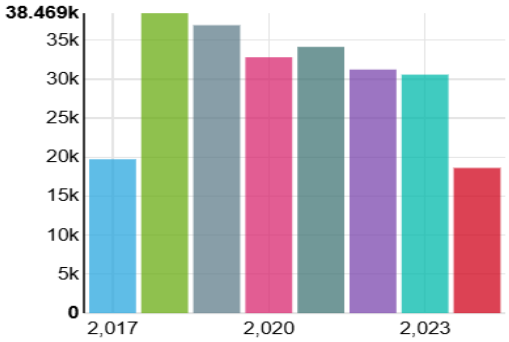

Fig. 1. Total Crashes Recorded in New Zealand (2017-2024) - Hue

The New Zealand Transport Agency's Crash Analysis System (CAS) is a popularly used data source capturing all police-reported crashes [4]. It provides mapping, trend analysis, and helps identify high-risk locations. CAS data has been widely used by both researchers and practitioners, and the system offers up-to-date and historical records dating back to 1 January 1991.

Fig. 2. The scene of the fatal crash near Mt Pleasant, south of Picton. (*Source: RNZ Trish Rawlings* [5])

Since 2011, several reviews have critically examined the methodologies and limitations of crash data analysis in New Zealand, with a focus on the application and integration of data from CAS. These reviews identify challenges and changing methods for effectively utilizing CAS data to understand crash risk, injury outcomes, and to inform policy interventions. Challenges have included variable data quality [4], under-reporting of some types of crashes (particularly minor or non-motorized road user crashes) [6] [7], changing police reporting standards over time, (including the way some types of road user are classified, e.g., cyclists under age 12 previously recorded as pedestrians) [8] [9], and regional inconsistencies in the way crashes are recorded. Methodological attention should be given to the analysis of CAS data to ensure that incomplete data, sampling bias, or unreported crashes, and integration difficulties when linking with health or hospital records are taken into account [10].

One of the key contributions is a 2024 systematic review and meta-analysis [10] that evaluated data linkage strategies employed to connect motor vehicle crash data, including CAS, with hospital and health outcome data. The review highlighted the importance of integrating multiple data sources to overcome the limitations in each of them, noting that linkage methodologies such as deterministic and probabilistic matching add significant value to the body of knowledge about the severity of injuries and contributory crash factors. The review also found gaps, especially in the

New Zealand context, where the integration of injury and crash data are underdeveloped. To complement this, a report on the topic of tourists crashes in Southern New Zealand [10] used a review approach to investigate the statistical methods used in analyzing CAS data coupled with sampling and reporting bias, which they found impacted the reliability and validity of analysis. Their report acknowledges intrinsic challenges including regional discrepancies and under-reporting of crashes, notably those involving vulnerable road users. It advocates for the adoption of supplementary datasets to assist statistical power and to overcome known shortcomings in CAS data.

Table 1 below presents a summary of peer-reviewed academic papers using New Zealand's Crash Analysis System (CAS) data, along with the method used, and key findings.

TABLE I.

Summaries of NZ Crash Data Studies		
Study / Source	***Methodology***	***Key Findings***
Hirsch, et al. (2021)	Analysis of 64 coded CAS variables; pedestrian crash data analysis	Detailed insights into pedestrian fatal and serious injury crashes; location and factor analysis
Walton, et al. (2020)	Spatial clustering of CAS crash data + speed modelling	Identification of crash hotspots related to speed; informed speed interventions
Smith, et al. (2020)	Linkage of CAS crash records with hospital trauma data	Combined crash and injury severity data for comprehensive analysis
NZ Transport Agency (2005)	Empirical Bayes, Poisson regression, multiple modelling methods	Framework for crash risk estimation and safety benefit calculations in NZ

A. Background and motivation

These studies show how CAS data has been explored in various traditional analysis and modelling techniques—ranging from statistical analysis, spatial modelling, to data linkage. While these have helped to drive crash risk understanding and improve road safety policies in New Zealand, it was unknown how effective the use of Big Data approaches may be to addressing some of the limitations in these traditional approaches to crash analysis. Thus, the present study sought to explore methodological enhancements using Big Data analysis to improve the robustness and applicability of crash risk models tailored to New Zealand's road safety landscape.

In New Zealand and the rest of the world, road trauma causes a large social and economic burden. Accurate measurement and timely analytics of road crashes therefore is essential in reducing harm. The New Zealand Crash Analysis System (CAS) is the main administrative source of police-reported crash events; however, it suffers known limitations (under-reporting, variable data quality, changing reporting standards and regional recording inconsistencies).

Connections between CAS and other datasets can make a significant contribution to improve inference about injury severity and contributory factors to crashes, however, linkage remains incomplete and is methodologically uneven throughout New Zealand. Meanwhile, innovations in Big Data engineering and scalable analytics (e.g., distributed processing, linkage algorithms, machine learning that is bias-aware) provide opportunities to address these constraints and generate more actionable, defensible policy evidence.

B. Research Objectives and Researcg Question

Understanding road crash patterns over time and the implications of severe weather, location, speed and other indicators related to severe incidents is an important area of study that is essential for improving government policy and potentially to saving lives. Thus, our research question was *Can a Big Data analysis approach reveal fresh insights in the CAS data that may not have been explored previously?*

As outlined in Figure 3. below, the research presented in this paper focused on applying Big Data techniques using Hadoop and MapReduce to analyse road vehicle crash data in New Zealand. The objectives of the study were to uncover patterns and trends related to crash incidents withing the CAS dataset, and to investigate them across different regions using geographic data. Additionally, the study aimed to enhance data processing and visualization through the application of Elasticsearch for indexing, with Kibana and PowerBI for visualization. The aim of the study was to uncover fresh insights from the data using Big Data techniques and, with that, offer policy and operational recommendations to improve the national response to vehicle crashes in New Zealand.

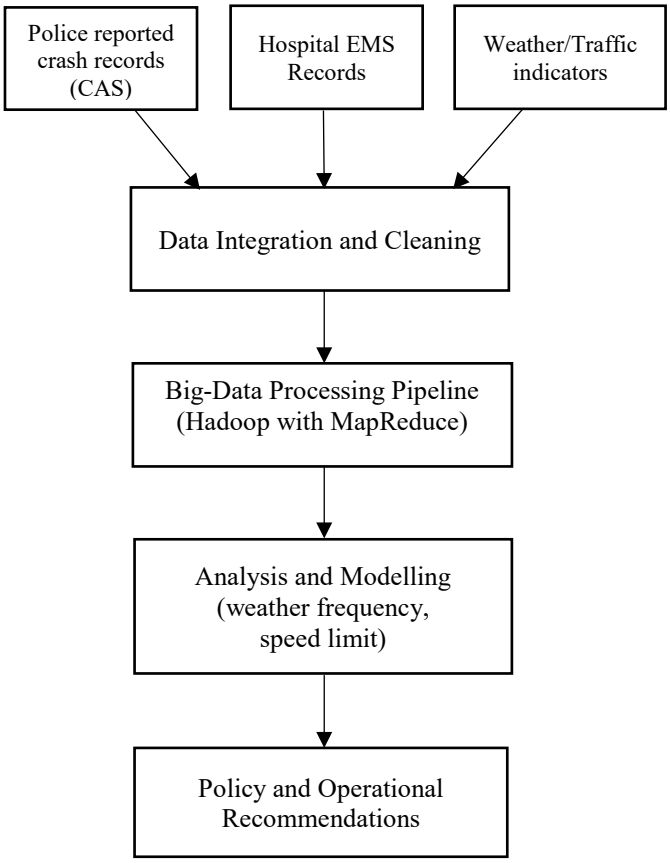

Fig. 3. Conceptual data and analysis pipeline

C. Key Themes for Analysis

To address the research question, two key themes were identified for analysis within the New Zealand Crash Analysis System (CAS) dataset. These themes represent important dimensions for crash risk that have been highlighted in previous studies but remain underexplored through a Big Data approach.

1) Weather Conditions and Crash Severity

The first theme of the present study was on examining the relationship between weather conditions and the severity of vehicle crashes. Weather has been widely recognised as a contributing factor in crash causation, where the road surface conditions, visibility, and driver behaviour have been affected [11]. However, as noted earlier, previous studies of CAS data have typically relied on traditional statistical approaches which might not have captured the non-linear or complex interactions between weather variables and the outcomes of crashes [12]. Using a distributed Big Data processing model based on Hadoop, Hive, and MapReduce, the present study investigated the association between the different weather conditions, including rain, snow, wind, and sleet, with crash severity levels.

2) Speed Limit and Crash Type

The second theme of the present study was on the effect of speed on crash types. Speed has consistently been identified as one of the most critical factors that can influence both the likelihood and the outcome of road accidents [13]. Traditional regression models often have assumed a uniform effect of speed across crash categories, potentially oversimplifying the role of speed [12]. Through the application of Big Data analytics with MapReduce, the present study disaggregates crash data based on the posted speed limits as well as crash contexts to reveal patterns that might not be identified by traditional statistical approaches.

D. Paper Structure

The research presented in the rest of this paper is organised into the following sections. Next, we present related research in the Literature Review, which summarizes existing studies which use Big Data techniques for crash data studies. Following this the Author Opinions are presented offering a perspective on the existing research. Then the Data and Statistical Analysis section discusses the data and statistical insights gleaned from the analysis of the CAS dataset. Following this, the Aggregation and Visualizations and the key findings presented. The MapReduce with Pseudo-code is then presented, explaining the logic and how it addresses the research problem. Finally, the paper concludes with a summary of findings and some directions for future research are offered.

II. Literature Review

Analysis to understand the factors that affect vehicle incidents have been an area of research focus over several decades. Yet in the absence of direct data that might identify the cause and effect relationships, most researchers have analysed crash phenomenon in terms of exploring the factors that may affect the frequency of crashes, such as, the number of crashes occurring in a specific geographical location over time, or to explore environmental conditions, such as weather incidents and elements affecting visibility [14]. Holiday periods when traffic out of urban areas peaks have also been a focus of study. Traditional statistical approaches and Bayesian methods have dominated the studies internationally [15] [16] and more recently we see a number of approaches using machine learning [17] [18] [13] [12] as well as emerging AI applications [20] [21]. Most existing studies have formulated models for road accident prediction as a regression or a classification problem [12]. Poisson regression and its variants are widely applied to model road accidents and severity [22]. However, there are fewer studies involving Big Data approaches to these areas of research interest.

It is important to point out that analysis of crash-frequency data does pose a range of challenges for researchers in terms of data characteristics. Characteristics such as low sample-means, under-reporting of crashes, time-variance explanatory variables, crash-type correlation, null-variables bias [23], and issues related to functional form and fixed parameters, such as assumptions about the relationship between explanatory variables and the dependent variable, or that the explanatory variable for crashes is to be identical for all observations [24]. Traditional models such as these that focus on incident count data often assume that the explanatory variables are directly influenced by the dependent variables in a linear manner [25] [24]. They also rely heavily on structured data. Road accident data is typically multi-variate in nature, but it is often modelled as univariate data in the predictive models published. However, an increasing number of studies indicate that non-linear functions provide a more accurate representation of the relationships between crash frequencies and their explanatory variables [26].

To tackle data-related challenges such as these, a number of innovative methodological approaches have been introduced seeking to improve the validity of findings. In comparison to the traditional approaches, Big Data analysis has created fresh opportunities for tackling the challenges of significantly large and varied data sets uncovering previously unknown patterns [16].

As an example of a Big Data approach, Chiou et al. [27] developed a genetic rule-mining model, which utilised a stepwise algorithm to identify patterns in crash data. The authors used a mixed logit model to examine the safety and risk variables related to severe crashes by including 29 derived rules. The results showed that seat belt use proved to be the most significant safety factor, while variables such as vehicle type, driver alcohol, and road surface conditions were the most significant risk contributors. Studies of this type show that a Big Data two-step mining technique is effective in investigating the risk factors in the severity of single-vehicle crashes.

In another study by Park et al., [28] the MapReduce technique was successfully applied to model a large-scale road accident dataset using k-means clustering and logistic regression (LR) for accident prediction. However, the study reported that the rate of false-positive predictions increased substantially as the number of features grew.

While no known Big Data approaches to date have been applied to New Zealand crash test data, the application of the Big Data and advanced data mining techniques, as exemplified in the Chiou et al. [27] study and the Park et al. [28], provide a foundation for using Big Data and MapReduce for identifying relationships in the New Zealand Crash Analysis System dataset.

Though the recent literature suggests an increasing interest in machine learning and Big Data methods in crash analysis,

important gaps still remain that justify the present study's Big Data research design.

First, most published literature continues to use traditional types of regression-based models (Poisson variants and Bayesian approaches) that assume a linear model, with fixed-form relationships and typically require well-structured, modest-sized data. These traditional regression-based approaches have difficulties with time-varying covariates, unobserved heterogeneity, under-reporting, and the high-dimensional, multivariate structure of contemporary crash datasets.

Second, although there are isolated Big Data demonstrations (e.g. with MapReduce enabled clustering or two stage mining), limited studies show how to scale end to end analyses, e.g., ingesting, cleaning, feature engineering, temporal aggregation, and model training on large national-level crash repositories, and none have applied such scalable pipelines to New Zealand crash data.

Third, prior MapReduce implementations have reported practical limitations (such as false-positive performance as feature counts grow) but do not explore how distributed preprocessing, partitioning, and SQL-like exploration can be used to address those issues or enable reproducible exploratory workflows.

Due to these reasons, the present study was based on the Hadoop MapReduce architecture using Hive. The distributed processing and input/output (I/O) efficiency of Hadoop allows for processing large data sets, such as the heterogeneous, and time-indexed crash data. Hive's SQL-like layer allows aggregations, partitioned temporal queries, and scalable extract-transform-load (ETL) processes. The MapReduce-based approach enabled dimensionality reduction and data pipelines that directly addressed the scalability and non-linearity gaps mentioned above. Through applying this Big Data stack to the New Zealand crash data, the study aimed to offer a novel methodological approach, enabling scalable discovery of non-linear and interaction effects that traditional statistical modelling may have missed.

III. Author Opinions

As mentioned above, most of the published studies followed traditional regression and parametric models relying on Poisson variants and Bayesian techniques, which are, perhaps, easy to interpret, are not designed to be applied in contemporary high-dimensional crash repositories. These traditional statistical models are commonly based on linearity and fixed-form assumptions, non-observation of time-varying covariates, have under-reporting biases, and collapse the multivariate structure into excessively simple univariate summaries.

Even the few Big-Data demonstrations in the literature involving vehicle crash data are apt to be biased. They show that MapReduce or clustering can scale computations, but cannot deliver end-to-end, reproducible pipelines combining preprocessing, bias correction, temporal partitioning and evaluation. The result is a gap between scalability and validity, which is arguably a weakness in the utility for findings to accurately inform policy.

While the present study does not extend into predictive or machine learning modelling, it provides methodological progress through its integration of Hadoop, Hive, and MapReduce to manage, aggregate, and analyse large-scale crash data efficiently. The Hadoop, Hive, and MapReduce architecture enabled an efficient processing of 242,540 records extracted from the CAS dataset. An analysis that would be computationally infeasible using conventional desktop-based tools.

Elasticsearch also improved the analytical process because it was able to index crash data geographically and Kibana and Power BI were able to give dynamic, visual interfaces to explore geographically and through time. The interactive map of New Zealand as the result of the analysis displays colour-coded crash frequencies by region (see Fig. 6 in section IV), and is an illustration of how Big Data infrastructure can transform our dataset into actionable, visual knowledge. Future extensions may be based on this framework either by adding a temporal trend analysis or by linking CAS data with additional datasets such as hospital records or weather or road condition sources to deepen causal understanding.

Big Data solves scale but is not free of bias and governance issues, which cannot be overlooked in the approach undertaken to the present study. Evidence from other research areas has demonstrated that naïve scaling only exaggerates existing biases (e.g., under-reporting of minor crashes [9]) and may lead to overconfident models if not calibrated. Thus, we argue for a future focus on bias correction that is able to capture and recapture estimates or introduce probabilistic linkages to the hospital data where possible to estimate under-reporting in the CAS dataset. A design approach as shown in Fig. 3 above.

We also suggest that more emphasis should be placed on provenance and reproducibility of Big Data studies where authors share HiveQL scripts, MapReduce job manifests and partitioned metadata in a publicly available a web-based platform such as GitHub and to expose them to reviewers. Privacy and ethics should also be kept in mind in such a scenario as even de-identified crash records can be re-identified when combined with spatial and temporal indices, so some elements of governance policies and minimisation rules should accompany any public-facing scripts or manifests as suggested.

The strongest contribution, however, of the Hadoop, Hive and MapReduce approach presented in the present study is operational. Through this approach, reproducible partitioned datasets can be generated that enable other researchers in the transport safety field to reuse models without necessarily re-running costly ingestion. Practically by producing a defined ETL recipe (Hive partitions by year, month, region; de-duplication rules; and rules to generate speed or weather-related features). Although our research might not take the field much beyond a single Big Data demonstration, it offers the basis of an approach to policy-relevant analytics which is scalable, transparent, and defensible.

IV. Dataset & Statistical Analysis

This project uses the below big data tools to achieve data storage, processing, analysis, and visualization. Hadoop handles and processes large data via distributed processing, whereas MapReduce divides analytical processes into smaller parallel ones to be more quickly performed. Hive and Hue offer a querying and visualization environment of the data on an SQL-like interface with a convenient environment. In the case of unstructured data, Elasticsearch can be used to perform fast text search and Kibana can be used to create interactive visualization such as heatmap and word cloud.

Finally, Power BI is used for achieving advanced graphic displays that are not supported by above tools.

The present study utilised data from the Crash Analysis System (CAS), which contains annual road crash statistics published by the New Zealand Ministry of Transport. The sample is based on Traffic Crash Reports filled in by New Zealand Police officers who attended fatal and injury crashes. The CAS database [29], also contains countrywide records of road accidents that took place as far back from Jan 1991 till today. The CAS also mentions in the original files that there is missing data for the years 2020 and 2021.

For the purposes of this study, we focused specifically on the crash data for selected from Kaggle [30], which had an original size of 243.7 MB. The usability of this file is rated 10/10, which represents the highest reliability. The reason the author did not choose the original file in CAS system is that it lacks month-related information, making it impossible to filter out the mixed data from 2024/2025. According to Kaggle author remarks, Dr. Maryam Rahmani was able to filter the data by the end of 2024.

The initial dataset contained 869,887 data entries that include all accidents in New Zealand between Jan/2000 and Dec/2024. This dataset comprises 54 variables which include different crash characteristics in which they include vehicle types, place of the crash, types of crashes, environmental factors, road features, driving behaviours, weather conditions etc. Due to the complexity of the original data, data cleaning and pre-processing involved two steps: removing irrelevant variables and filtering out old years' records. The data cleaning process also involved removing all duplicate entries, irrelevant values, and missing data from the initial dataset.

After the cleaning process, the final dataset (see Fig. 4) was reduced to 19.2 MB and filtered to include data from January 2017 to December 2024, containing 242,540 road accidents. This project use these key variables of crashYear, crashSeverity (vehicle damage condition), directionRoleDescription (cardinal directions), tlaName (region where the crash happened), speedLimit, WeatherA and WeatherB (weather conditions when crash happened).

The second dataset is a shapefile downloaded in Stats NZ [31], based on the WGS84 (EPSG:4326) coordinate system. Then, it was manipulated in Python (GeoPandas Code) to get latitude and longitude coordinates, as presented in Figure 5. There were small adjustments of some regional boundaries to rectify small geographic errors.

A copy of both two datasets were submitted along with this report.

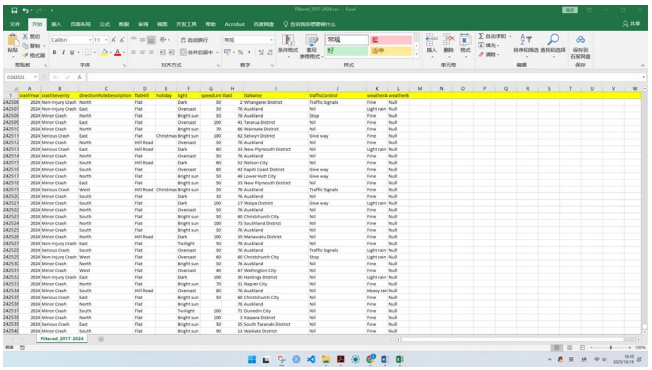

Fig. 4. New Zealand Traffic Crash Filtered Data from Year 2017 to 2024

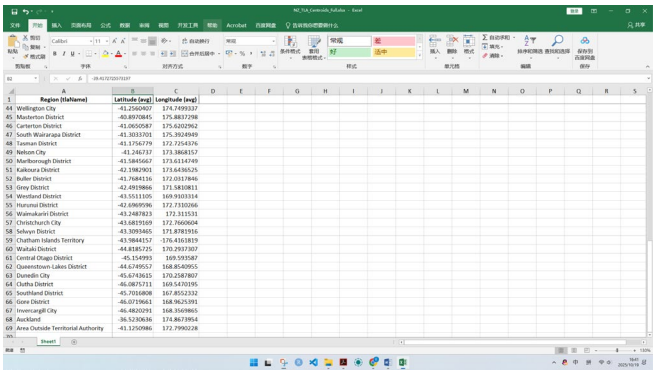

Fig. 5. New Zealand Regions Latitude & Logitude

```
# Python Geopandas Code
    import geopandas as gpd
    import pandas as pd

# Read Stats NZ ShapeFile
    boundary = gpd.read_file("territorial-authority-
2023-clipped-generalised.shp")

# Convert WGS84 into lag/log
    boundary = boundary.to_crs(epsg=4326)

# Calculate an average spot for each region
    boundary["centroid"] = boundary.geometry.centroid
    boundary["Longitude"] = boundary.centroid.x
    boundary["Latitude"] = boundary.centroid.y

# Select Existing Column Region + lat + lon
    out = boundary[["TA2023_V1_00_NAME",
"Latitude", "Longitude"]].copy()
    out = out.rename(columns={
        "TA2023_V1_00_NAME": "Region (tlaName)",
        "Latitude": "Latitude (avg)",
        "Longitude": "Longitude (avg)"
    })

# Output Excel
    out.to_excel("NZ_TLA_Centroids_Full.xlsx",
index=False)

print("Complete! NZ_TLA_Centroids_Full.xlsx")
```

For an overview of geographic crash distribution in New Zealand, the author imports the structured data into Elastic Kibana and generate the geographic view in Figure 6. The diameter and color of each circle on the map are based on the total number of crashes in a given area. In particular, light green shows low crash volumes, green represents medium levels, yellow shows high crash volumes (Chrischurch), while red shows the biggest and hottest areas of crashes, such as Auckland. One interesting observation is there are duplicate spots in some nearby districts in Wellington City.

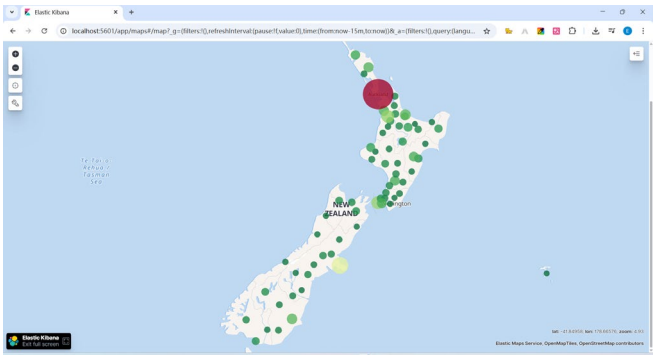

Fig. 6. New Zealand Crash Geographic View (2017-2024) Visualised in Elastic Kibana - Cluster Map

Figure 7 below shows the total road crashes in the New Zealand regions, which are created in Hive with the SQL query. Below Figures 8 and 9 were created in Power BI to provide a clearer interface, while the data was queried by HQL.

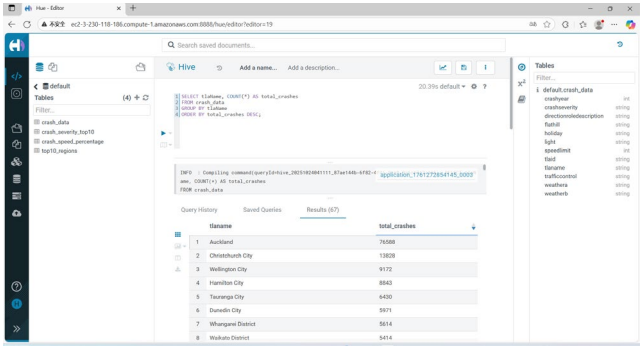

Fig. 7. Total Crash Overview from 2017 to 2024 Descending Order - Hue

Figure 8 shows the top 10 regions in New Zealand of the highest number of road crashes between 2017 and 2024 based on the level of the severity of the crash. In the chart legend, each road accident was categorised into one of four classes of severity: fatal, serious, minor, and non-injury [4]. These can be defined as:

- Fatal Crash: A road crash that caused deadly results.
- Serious Crash: A road crash that results in third-party medical treatment and hospitalisation.
- Minor Crash: A road crash in which people might experience bruising or small cuts but do not require medical treatment.
- Non-injury Crash: A road accident with no one being hurt

When we deep dive into the region's category in Figure 8, things become interesting. The highest crash severity for all regions is Non-Injury Crash, then followed by Minor Crash. Auckland is dramatically different among top 10 regions, with the crash frequency being nearly 5 times greater than in the other locations.

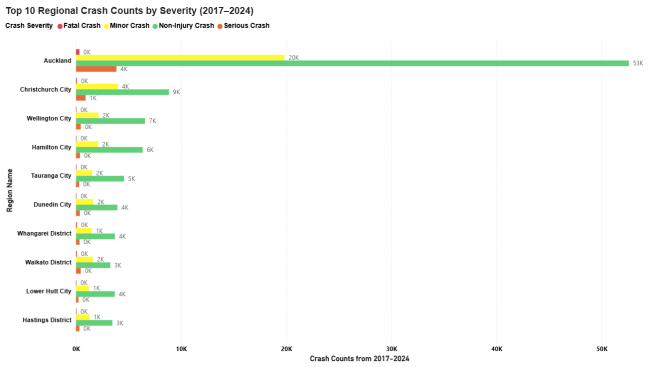

Fig. 8. Total Crash by Top 10 Regions Visualised in Power BI - Bar Chart

Figure 9 shows the annual crash trends of Auckland, Christchurch and Wellington. Between 2017 and 2018, total crash numbers increase to the peak level in all three regions. A small decrease was observed after this, then came a stable period from the years 2019 to 2023. Finally, Auckland shows a significant decline by the end of 2024, whereas Christchurch and Wellington show only a slight drop by the end.

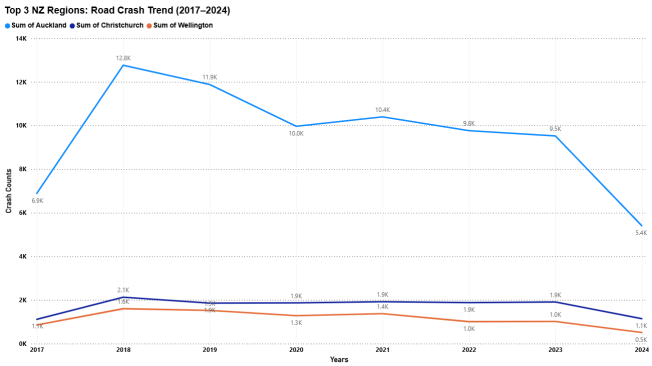

Fig. 9. Top 3 New Zealand Regions Crash Trends (2017-2024) Visualised in Power BI - Line Chart

V. Aggregation and Visualisations

In this section, we perform the advanced analysis of road crash trends in New Zealand in 2017-2024, with respect to crash severity level, weather conditions, and influence of speed limits.

The Figure 10 word cloud shows the combined impact of primary (Weather A) and secondary (Weather B). We can see that Fine_Null weather appears the most, then Light rain_Null comes with second level, Fine_Strong Wind, Heavy rain_Nan and Snow_Strong Wind come with thrid position. Compared with figure 11, we came up with the conclusion that extreme weather crashes happen less often than expected. Although a strong wind usually seems to be a secondary cause of crashes, it is not always a key cause of the severity of crashes.

Fig. 10. Weather Conditions Associated with Road Crashes Visualized in Elastic Kibana - Word Cloud

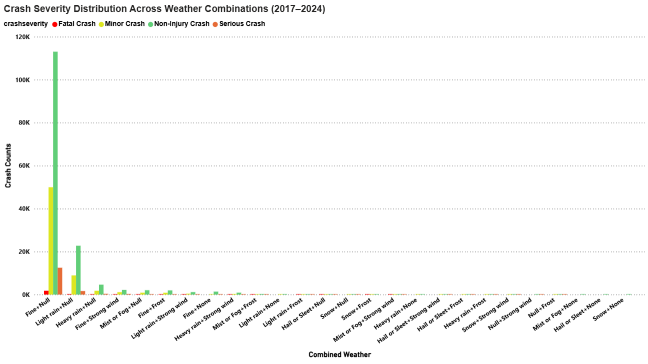

Fig. 11. Weather Conditions Associated with Crash Severity Visualized in Power BI - Bar Chart

Then we consider if the speed limit would be the main reason caused the crashes. As shown in Fig. 12, most crashes occurred in areas with a speed limit between 40-60 km/h (54.94%), followed by roads with speed limits of 100 km/h or more (28.01%). Regarding crash severity, Figure 13 shows that the majority were non-injury crashes (66.01%), with minor crashes (26.66%), serious crashes (6.4%), and fatal crashes (0.94%). Figure 14 clearly shows the severity levels rise with speed, which demonstrates that while low-speed areas have more crash counts, high-speed crashes tend to be deadlier. In conclusion, although high-speed regions are a major cause of deadly accidents, they are not the decisive factors.

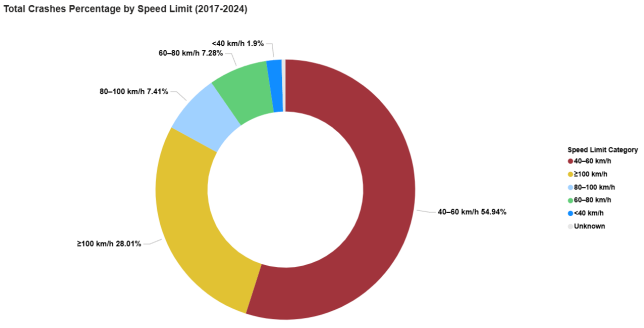

Fig. 12. Speed Limit Conditions Associated with Road Crashes Visualised in Power BI - Donut Chart

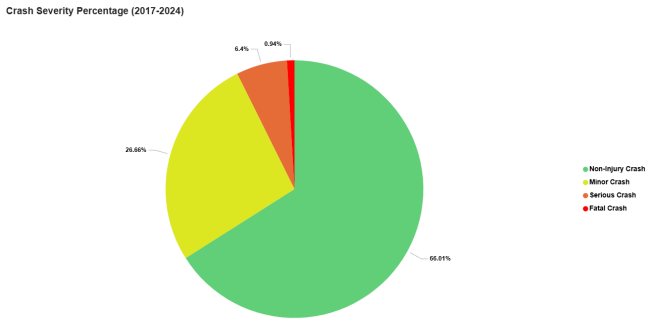

Fig. 13. Crash Severity (2017–2024) Visualised in Power BI - Pie Chart

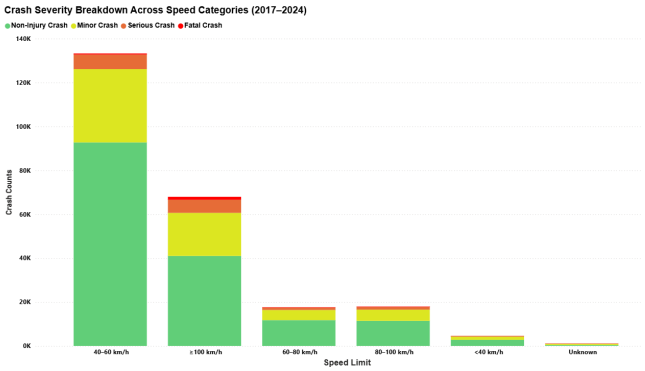

Fig. 14. Crash Severity by Speed Category (2017–2024) Visualised in Power BI - Bar Chart

VI. MapReduce Pseudo-code with Explanations

MapReduce is the distributed model of computing, which is developed to work with large datasets. It may be applied to both structured (such as the use of tables) and unstructured (such as text or logs) datasets [32]. It divides complex computing tasks into small manageable operations that can be executed in parallel on a number of machines. The architecture enables systems to support a large amount of data work load efficiently without compromising fault tolerance or scale, hence it is perfect in tasks such as search indexing, log analysis or even recommendation engines.

The MapReduce workflows comes with below four key steps:

- **Split:** The input data is broken down into smaller units.
- **Mapper:** Every chunk is run to produce key-value pairs.
- **Shuffle and Sort:** The system sorts and group intermediate data using key.
- **Reducer:** Aggregated outputs are produced per key.

Python is a popular language for running MapReduce due to its readability, ease with which it can be used to test ideas, and the availability of helpful libraries. Unlike Java, which is commonly used in Hadoop, Python allows the development of Mapper and Reducer as normal scripts using tools such as Hadoop Streaming or mrjob [33]. It is suitable for medium-sized projects and is great for academic projects because its code is so clear and straightforward.

Step 1: According to the crash records that have been gathered between 2017 and 2024, the following Python MapReduce code will count the total amount of crashes that took place in each region (tlaName).

To begin with, we import two MapReduce libraries: MRJob and MRStep. Then we define the two-step function inside the class CrashCountByRegion.

In the Map phase we scan through the document and use commas to separate cells. Meanwhile, we skip the empty region value and only select the value in the field 8 that is "tlaName", which represents the region where the crash took place. For these valid records, we yield the region as the key value with 1.

In the Reduce step, we sort out the same regions and count all the values where the regions match. Lastly, main function prints out the values.

#Step 1: Overview of Crashes from 2017 to 2024 by Region (Fig. 7).

```
# Import API
    from mrjob.job import MRJob
    from mrjob.step import MRStep

    class CrashCountByRegion(MRJob):

# Define two steps
        def steps(self):
            return [
                MRStep(mapper=self.mapper_get_region,
                       reducer=self.reducer_count_region)
            ]

        def mapper_get_region(self, _, line):

# Split words with comma
# Confirm filed 8 exist
            fields = line.split(",")
            if len(fields) > 8:
                region = fields[8].strip()

# Skip empty regions
                if region and region != "tlaName":
                    yield region, 1

# Expected output: ("Auckland", 1)

        def reducer_count_region(self, region, counts):
            yield region, sum(counts)

    if __name__ == '__main__':
  CrashCountByRegion.run()

# Expected output: "Auckland" 765888
#                  "Christchurch City" 13828
```

Step 2: This MapReduce code uses the combination of the weather attributes (weatherA and weatherB) to determine the relationship between the two weather conditions and occurrence of crashes.

During the Map phase, we must filter off the Null value first. However the original file has numerous nulls in the column WeatherB and this column cannot be filtered individually. Instead, we combine WeatherA and WeatherB together (e.g., Fine strong wind) and filter out where both values are empty. After filtering, we publish the key with the value of 1.

In the Reduce stage, all same values of weather combinations are added together to determine which one is the most frequently found conditions in accidents.

#Step 2: Weather Conditions Associated with Road Crashes (Fig. 10).

```
    class WeatherComboAnalysis(MRJob):
        def steps(self):
            return [
             MRStep(mapper=self.mapper_weather_combo,
                    reducer=self.reducer_combo_count)
            ]

        def mapper_weather_combo(self, _, line):
            fields = line.split(",")
            if len(fields) > 11:
                weatherA = fields[10].strip().lower()
                weatherB = fields[11].strip().lower()

# Clean both Null value
                if weatherA != "null" and weatherB != "null":
                    combo = f"{weatherA}_{weatherB}"
                    yield combo, 1

# Expected output: ("fine_null", 1)
#                  ("mist_null", 1)
#                  ("light rain_strong wind", 1)

        def reducer_combo_count(self, combo, counts):
            yield combo, sum(counts)

    if __name__ == '__main__':
        WeatherComboAnalysis.run()

# Expected output: ("fine_null", 37895)
#          ("light rain_null", 8025)
#          ("light rain_strong wind", 2354)
```

Step 3: This MapReduce code explains the different speed limits influenced the number of the crashes.

At the Map phase, the speedLimit of every crash record is divided into five categories (<40, 40-60, 60-80, 80-100, >100 km/h). The key of each category is the value 1. And when there is NULL or can not be read then the reducer assigns the entry to an "Unknown" category. This ensures that the job proceeds without missing incomplete data and final results are accurate.

During the Reduce stage, the counts of the same range of speed are added to give the total number of crashes per group of speed limits.

#Step 3: Speed Limit Conditions Associated with Road Crashes (Fig. 12).

```
    class SpeedLimitGroupAnalysis(MRJob):
        def steps(self):
            return [
                MRStep(mapper=self.mapper_group_speed,
                       reducer=self.reducer_count_speed)
            ]

        def mapper_group_speed(self, _, line):
            fields = line.split(",")
            if len(fields) > 6:

# Define 5+1 groups
                try:
                    speed = float(fields[6])
                    if speed < 40:
                        category = "<40 km/h"
                    elif 40 <= speed < 60:
                        category = "40-60 km/h"
                    elif 60 <= speed < 80:
                        category = "60-80 km/h"
                    elif 80 <= speed < 100:
                        category = "80-100 km/h"
                    else:
                        category = ">=100 km/h"
                    yield category, 1

# Store Null Value
                except:
                    yield "Unknown", 1

# Expected output: ("40-60 km/h", 1)

        def reducer_count_speed(self, category, counts):
            yield category, sum(counts)

    if __name__ == '__main__':
        SpeedLimitGroupAnalysis.run()

# Expected output: "<40 km/h"      245
#                  "40-60 km/h"    124735
#                  "60-80 km/h"    41372
```

VII. Conclusion & Future Issues

In conclusion, this paper illustrates that weather conditions and the speed limits set on the road are correlated with road accidents in New Zealand, they are not necessarily the major causes of road accidents. As an example, visual reports indicated that most of the crashes took place despite the fine weather conditions and speed range of 40-60 km/h, which are conditions normally regarded as safe. Nevertheless, secondary weather factors such as high wind or incidences of excessive speeds more than 100 km/h became a contributory factor indicating that isolated conditions cannot be used as a sole cause of crashing.

Based on the limitations of this study, the following directions for future work are suggested:

Weather Elements: Future analysis can include a third variable such as time of day to investigate if crashes are more likely to occur at night, midday, or rush hour, not only in different weather conditions This could help to reveal new patterns of accident risk.

Speed Limit Group 40-60 km/h: This group had the largest amount of crashes, therefore it requires more studies of driver behaviour. For example, are crashes in this range caused by aggressive driving, such as failure to give way, or is caused by external environmental factors such as glare from the sun or visibility of road?

Auckland Region: During this project study period from 2017 to 2024, Auckland had the largest number of road crashes of any region in New Zealand. So future studies could focus on a specific suburb or street and investigate what elements cause this to happen.

More importantly, this study indicates that the MapReduce framework plays a critical role in managing and analysing massive crash data enhance transportation safety. Future research could also test machine learning (including bias-aware models) for crash prediction or connect different data sources (e.g. crash records and hospital reports) to provide more reliable information on the severity of crashes.

Recurrent Neural Networks (RNNs) have been shown to be effective in the analysis of time-series and sequential data and are frequently applied in prediction and classification [34]. LSTM and GRU are advanced types of the RNN model that are effective in working with long-term patterns and length sequences. This is why they come in handy in fields such as healthcare, finance, IoT, and data center management [35] [36].

PyTorch (Python API) is a deep learning framework which enables a more convenient way of creating and testing models, which is used to analyze RNNs. This loose organization suits agile development and cloud-based processes, in situations where the model has variable size and structure [37].

VIII. Acknowledgment

The authors acknowledge and thank Dr Anuradha Singh from the School of Engineering, Computer and Mathematical Sciences at the Auckland University of Technology for ger guidance in the preparation of this study and paper.

IX. References

[1] J. Millar, Retrospective analysis of fatal car crashes using near miss forecasting, Christchurch: Unpublished MURR thesis, University of Canterbury, NZ, 2023.

[2] Ministry of Transport, “Domestic Transport Costs and Charges Study: Working Paper D1 Costs of Road Traffic Accidents,” Ministry of Transport, 2023.

[3] Ministry of Transport, “Safety - annual statistics,” Ministry of Transport, 2024.

[4] New Zealand Transport Agency, "Crash Analysis System (CAS)," 2025. [Online]. Available: https://www.nzta.govt.nz/safety/partners/crash-analysis-system/

[5] Radio New Zealand, "Seven killed in head-on crash south of Picton," 19 06 2022. [Online]. Available: https://www.rnz.co.nz/news/national/469392/seven-killed-in-head-on-crash-south-of-picton

[6] A. Smith, J. Garvitch, K. Clark and G. Christey, "Police motorcycle crash casualty reports and their linkage with hospital trauma admissions in the Midland Region of New Zealand," *Journal of Road Safety,* vol. 31, no. 2, pp. 13-22, 2020.

[7] H. H. Jama, R. H. Grzebieta, R. Friswell and A. S. McIntosh, "Characteristics of fatal motorcycle crashes into roadside safety barriers in Australia and New Zealand," *Accident Analysis and Prevention,* vol. 1, no. 43, pp. 652-660, 2011.

[8] L. Hirsch, H. Mackie and I. McAuley, "Fatal footsteps: understanding the Safe System context behind New Zealand's pedestrian road trauma.," *Journal of Road Safety,* vol. 32, no. 1, pp. 5-16, 2021.

[9] New Zealand Transport Agency, "Review of Crash Analysis System (CAS) data quality and under-reporting issues: Impacts on crash risk modelling," New Zealand Transport Agency, 2022.

[10] S. Karimi, A. Hosseinzadeh, R. Kluger, T. Wang, R. Souleyrette and E. Harding, "A systematic review and meta-analysis of data linkage between motor vehicle crash and hospital-based datasets.," *Accident Analysis & Prevention,* vol. 197, p. 107461, 2024.

[11] M. Asgarzadeh, D. Fischer, S. K. Verma and T. K. Cour, "The impact of weather, road surface, time-of-day, and light conditions on severity of bicycle-motor vehicle crash injuries," *American Journal of Industrial Medicine,* vol. 61, no. 7, pp. 556-565, 2018.

[12] S. Ahmed, M. A. Hossain, M. M. I. Bhuiyan and S. K. Ray, "A comparative study of machine learning algorithms to predict road accident severity," in *20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS)*, 390-397, 2021.

[13] C. Jurewicz, A. Sobhani, J. Woolley, J. Dutschke and B. Corben, "Exploration of vehicle impact speed–injury severity relationships for application in safer road design," *Transportation research procedia,* vol. 14, pp. 4247-4256, 2016.

[14] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives," *Transportation Research Part A: Policy and Practice,* vol. 44, no. 5, pp. 291-305, 2010.

[15] S. Anowar, S. Yasmin and R. Tay, "Comparison of crashes during public holidays and regular weekends.," *Accident Analysis & Prevention,* vol. 15, pp. 93-97, 2013.

[16] L. Skaug, M. Nojoumian, N. Dang and A. Yap, "Road Crash Analysis and Modeling: A Systematic Review of Methods, Data, and Emerging Technologies," *Applied Sciences,* vol. 15, no. 13, p. 7115, 2025.

[17] D. Santos, J. Saias, P. Quaresma and V. Beires-Nogueira, "Machine learning approaches to traffic accident analysis and hotspot prediction," *Computers,* vol. 10, no. 12, p. 157, 2021.

[18] S. Pourroostaei-Ardakani, X. Liang, K. Mengistu, R. S. So, X. Wei, B. He and A. Cheshmehzangi, "Road car accident prediction using a machine-learning-enabled data analysis," *Sustainability,* vol. 15, no. 7, p. 5939, 2023.

[19] X. Wen, Y. Xie, L. Jiang, Y. Li and T. Ge, "On the interpretability of machine learning methods in crash frequency modeling and crash modification factor development," *Accident Analysis & Prevention,* vol. 168, p. 106617, 2022.

[20] T. Huang, S. Wang and A. Sharma, "Highway crash detection and risk estimation using deep learning.," *Accident Analysis & Prevention,* vol. 135, p. 105392, 2020.

[21] Z. Zhang, Q. Nie, J. Liu, A. Hainen and N. Islam, "Machine learning based real-time prediction of freeway crash risk using crowdsourced probe vehicle data," *Journal of Intelligent Transportation Systems,* vol. 28, no. 1, pp. 84-102, 2024.

[22] A. Abdulhafedh, "Crash frequency analysis," *Journal of Transportation Technologies,* vol. 6, no. 4, p. 169, 2016.

[23] N. K. ChikkaKrishna and P. Manoranjan, "Identifying safety factors associated with crash frequency and severity on nonurban four-lane highway stretch in India," *Journal of Transportation Safety & Security,* vol. 9, no. 1, p. 6–32, 2016.

[24] D. Lord, A. Manar and A. Vizioli, "Modeling crash-flow-density and crash-flow-v/c ratio for rural and urban freeway segments," *Accident Analysis and Prevention,* vol. 37, no. 1, pp. 185-199, 2005.

[25] S. P. Miaou, J. J. Song and B. K. Mallick, "Roadway traffic crash mapping: a space-time modeling approach," *Journal of Transportation and Statistics,* vol. 6, no. 1, pp. 33-57, 2003.

[26] J. A. Bonneson and M. P. Pratt, "Procedure for developing accident modification factors from cross-sectional data," *Transportation Research Record,* no. 2083, pp. 40-48, 2005.

[27] Y. C. Chiou, L. W. Lan and W. P. Chen, "A two-stage mining framework to explore key risk conditions on one-vehicle crash severity," *Accident Analysis and Prevention,* vol. 50, pp. 405-415, 2013.

[28] S. H. Park, S. M. Kim and Y. G. Ha, "Highway traffic accident prediction using VDS big data analysis," *The Journal of Supercomputing,* vol. 72, no. 7, pp. 2815-2831, 2016.

[29] Waka Kotahi NZ Transport Agency, "Crash Analysis System (CAS) data," 2024. [Online]. Available: https://opendata-nzta.opendata.arcgis.com/datasets/NZTA::crash-analysis-system-cas-data-1/about

[30] M. Rahmani, "Crash Analysis System (CAS) Data — New Zealand Traffic Crash Dataset," 2024. [Online]. Available: M. Rahmani, "Crash Analysis System (CAS) Data — New Zealand Traffic Crash Dataset," 2024. [Online]. Available: https://www.kaggle.com/datasets/maryamrahmani

[31] Statistics New Zealand, "Territorial authority 2023 (generalised)," 2022. [Online]. Available: https://datafinder.stats.govt.nz/layer/111194-territorial-authority-2023-generalised/

[32] F. Li, B. C. Ooi, M. T. Özsu and S. Wu, "Distributed data management using MapReduce," *ACM Computing Surveys,* vol. 43, no. 3, pp. 1-42, 2014.

[33] S. Leo and G. Zanetti, "Pydoop: a Python MapReduce and HDFS API for Hadoop," in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing, HPDC 2010*, Chicago, 2010.

[34] F. Deng and L. Zhang, "Performance optimization and interpretability of recurrent Sigma-Pi-Sigma neural networks on application of IoE data," *IEEE Internet Things Journal,* vol. 12, pp. 3639-3653, 2025.

[35] Q. Deng, "Applying recurrent neural networks to time-series analysis in big data for decision support," in *2024 IEEE 4th International Conference on Data Science and Computer Application (ICDSCA 2024)*, Dailian, 2024.

[36] M. Malin and J. Suutala, "Data center resource usage forecasting with convolutional recurrent neural networks," in *The Linköping Electronic Conference Proceedings* , Linköping, 2025.

[37] O. Novac, M. Chirodea, C. Novac, N. Bizon and M. Opro, "Analysis of the application efficiency of TensorFlow and PyTorch in convolutional neural network," *Sensors,* vol. 22, p. 8872, 2022.