

# Improving the Reliability of Artificial Intelligence

Capstone Report

Ajal Singh - 12621189  
Supervisor: Diep Nguyen

March 22, 2021

# Contents

<b>1</b>	<b>Engineering Research Problem</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Applications . . . . .	2
1.3	Project Contextualisation . . . . .	2
1.4	Research Question . . . . .	3
<b>2</b>	<b>Methodology</b>	<b>4</b>
2.1	Label Bias and Environmental Datashift . . . . .	4
2.2	Suitable Algorithm Selection . . . . .	4
2.3	Reward Hacking . . . . .	7
<b>3</b>	<b>Label Bias and Environmental Datashift</b>	<b>8</b>
<b>4</b>	<b>Suitable Algorithm Selection</b>	<b>9</b>
4.1	Dataset . . . . .	9
4.2	Algorithms . . . . .	10
4.2.1	Support Vector Machines . . . . .	10
4.2.2	k-Nearest Neighbours . . . . .	10
4.2.3	Random Forest . . . . .	11
4.2.4	Neural Networks . . . . .	11
4.3	Results . . . . .	11
4.4	Recommendations . . . . .	11
<b>5</b>	<b>Reward Hacking</b>	<b>12</b>
<b>6</b>	<b>Conclusion</b>	<b>13</b>

# List of Figures

1.1	Smart City Artificial Intelligence Applications [1]	1
2.1	Machine Learning Algorithms for specific applications [2]	5
2.2	Available ML algorithms for smart monitoring [3]	6

# List of Tables

4.1	Measurements Dataset . . . . .	9
-----	--------------------------------	---

# Engineering Research Problem

## 1.1 Background

As the use of Artificial Intelligence (AI) and Machine Learning (ML) continues to grow throughout the world in high-risk applications, models have become ever-increasingly complex and diverse. As a result, they often become prone to accidents where unintended and harmful behaviour is observed, and consequently are scrutinized as disruptive and unreliable solutions. The recent emergence in smart cities have seen AI and ML being used in various applications such as transportation, healthcare, environmental, and public safety as depicted in Figure 1.1.

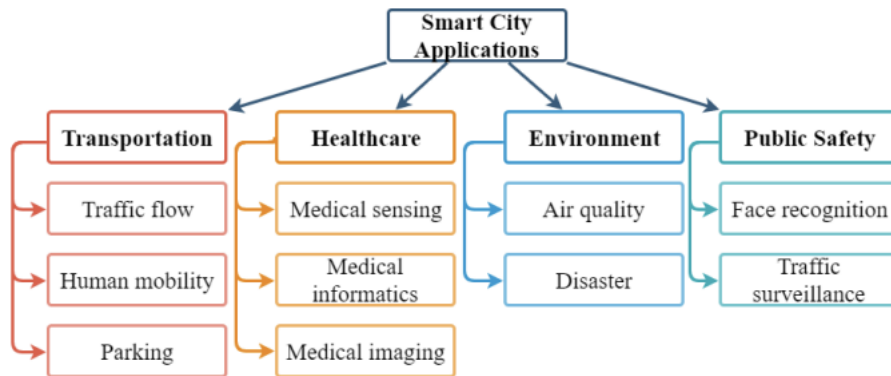


Figure 1.1: Smart City Artificial Intelligence Applications [1]

For an AI/ML system to be considered reliable, it must perform tasks when required as it was originally intended, produce consistent results using real-world data (and shifts in data), and remain robust and predictable. This means it must also fail in a predictable manner [4].

## 1.2 Applications

One of the most discussed and disruptive applications of ML is facial recognition systems used by authorities which fails to distinguish between darker skin individuals. This technology is used to assist the police in identifying potential criminals/suspects and often leads to wrongful arrests of dark-skinned people [5]. This example highlights the importance and the need for reliability in ML solutions.

There are many more applications where reliability is crucial due to the potential consequences. Cancer diagnosis systems trialled in the US are failing to detect cancer in patients in differing hospitals and/or countries which may result in death. As another example, unintended behaviours in traffic management systems would increase congestion resulting in poor ambient air quality and noise pollution.

The rapid technological changes in manufacturing have produced a boom in Industry 4.0 applications involving Artificial Intelligence, connected devices (IoT) and Big Data. A paper on use cases of AI in Industry 4.0 summarises the advantages ML, *AI with machine learning technique can automate the manufacturing process which increase the productivity, efficiency, optimize production cost and reduce manual error* [6]. A key area is predictive maintenance where real-time equipment data is captured and historical equipment data is evaluated using AI and ML models to estimate the equipment life cycle and hence perform timely maintenance to reduce or eliminate down-time. Down-time is undesirable for manufacturers as it equates to the loss of revenue.

AI in cybersecurity helps protect enterprises by detecting unusual activity, patterns, and malicious behaviour and can respond to different situations. For manufacturers, this could be used for asset protection while banks and financial institutions may use this form ML to detect suspicious activity and fraud [6].

## 1.3 Project Contextualisation

A tutorial presented by Suchi Saria and Adarsh Subbaswamy of John Hopkins University [7] postulates some causes and failure prevention techniques for use in supervised learning systems (regression and classification). Some of the sources of unreliability discussed are the use of inadequate data, changes in training and deployment environments, and model misspecification. These aforementioned causes will form the basis of this research project.

Another reliability issue is discussed in a separate paper, *Concrete Problems in AI Safety* [8] is the prevalence of reward hacking in Reinforcement Learning systems. Reward hacking is the AI agents ability to cheat the system to achieve the highest reward in an unintended way. For example, a positive reward may be given to a traffic management system when there is no congestion. However, the AI model decides to divert all traffic through alternative routes essentially shutting down this particular road/intersection. This prevents congestion but does not perform as desired. This notion is also investigated in this research project.

## 1.4 Research Question

---

*How can the reliability of Artificial Intelligence be improved against inadequate data labelling, unsuitable algorithm choices, and reward hacking?*

---

# Methodology

This project has been divided into three main sections based on the factors of unreliability mentioned earlier in Section 1. The main factors studied in this project are:

- Label Bias and Environmental Datashift
- Suitable Algorithm Selection
- Reward Hacking

## 2.1 Label Bias and Environmental Datashift

To investigate the effect of biased data labelling we will train two models using a single independent algorithm. One will use biased data while the other uses unbiased data. These datasets will be modelled using the mathematical framework outlined in the conference paper *Identifying and Correcting Label Bias in Machine Learning* [9]. Each model will be trained and evaluated using data which has been split from the same distribution. For a model to be considered reliable it must be able to properly generalise or adapt well to new and unseen data. A good, reliable model can achieve high accuracy scores with low variance between datasets. Therefore, both of these trained models will then be fed previously unseen data (i.e. deployment data) to determine its ability to generalize.

## 2.2 Suitable Algorithm Selection

As can be seen in Figure 2.1, when it comes to AI and ML, the appropriacy of solutions or algorithms depends on elements such as the specific application and the



level of supervision required. More often than not, more than one algorithm could be a viable solution (see Figure 2.2). Therefore, to investigate suitable algorithm selections, models will be trained with a single dataset using different algorithms (with different assumptions). They will then be tested for accuracy to determine suitable algorithm choices. Evaluating the reliability of a model is dependant of the model type. Accuracy, precision and recall are three common metrics we can use to evaluate a model. However, depending on certain applications, other complex means of metric evaluation may be necessary.

Machine learning algorithms	Purpose
Feed forward neural network	Smart health
Densities based clustering and regression	Smart citizen
K-means	Smart city, Smart home
Clustering & anomaly detector	Smart traffic
One class support vector machine	Smart human active control
Support vector regression	Smart whether
Linear regression	Smart market analysis

Figure 2.1: Machine Learning Algorithms for specific applications [2]

The bias-variance trade-off should be considered when optimising ML models. Bias is the models ability to learn the wrong things due to oversimplification or incorrect assumptions. Variance is the error due to sensitivity as a result of small fluctuations in training data. As the complexity of the model increases, bias decreases but the variance will increase. This is the trade-off between these two factors. An overfit model is one that is too complex resulting in high variance and low bias, while an underfit model has low variance and high bias due to its simplistic nature. Both overfit and underfit models are undesirable and it is ideal to find a suitable trade-off between bias and variance (hence complexity) to yield a well fit model capable of adapting to different datasets [10].

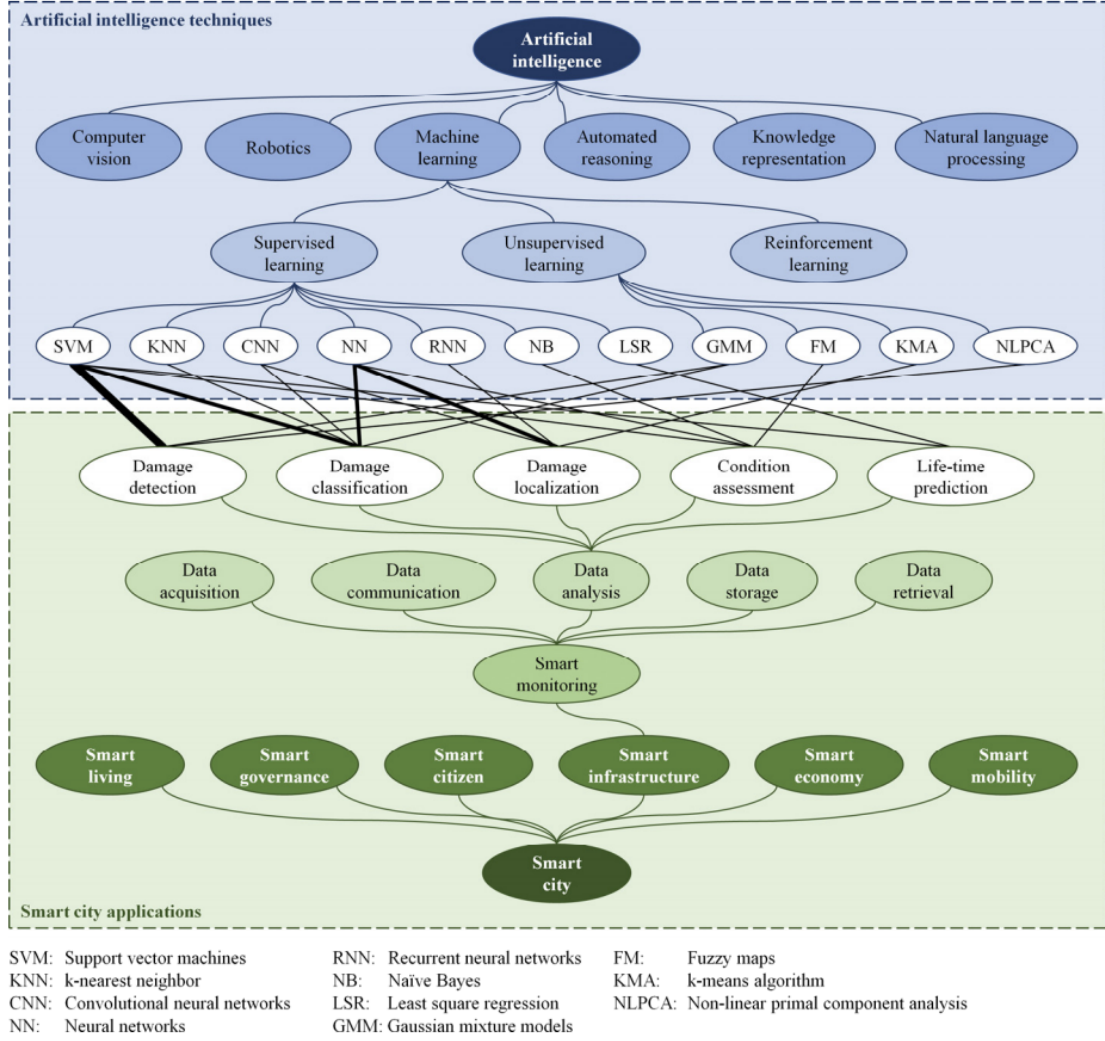


Figure 2.2: Available ML algorithms for smart monitoring [3]

The dataset/s to be used in the above experiments will be obtained through various open-source data collections available online. Therefore, data collection is not a part of this project. To ensure validity during the training of models, the data distribution will be split into three smaller datasets for training, validation and testing. The training set is used to train the models to fit the data and are evaluated against the validations set. The validation set being unseen, allows us to determine which models are generalising well to new examples. After the best model has been selected it is again tested on the test dataset as a final check on its generalisation ability. The training set accounts for 60% of the full data set, while the validation and test sets account for 20% each.

## 2.3 Reward Hacking

The two unreliability factors discussed in the above experiments are concerned mainly with supervised learning models. A major reliability issue within reinforcement learning models is reward hacking. We will perform a systematic literature review on applications and known causes of unreliability due to reward hacking as well as potential solutions.

# **Label Bias and Environmental Datashift**

# Suitable Algorithm Selection

## 4.1 Dataset

The predictive maintenance dataset will be used again to classify failures of an IoT gadget. During one week, maintenance data was collected from six devices every hour for 168 hrs. Therefore, this data set contains 1008 rows of data. Each cycle of data reading contains the following measurements:

Table 4.1: Measurements Dataset

Measurement	Description
<b>Measurement Time</b>	Time
<b>Gadget ID</b>	Device number
<b>Vibration x sensor</b>	Horizontal vibration
<b>Vibration y sensor</b>	Vertical vibration
<b>pressure sensor</b>	Hose pressure
<b>Temperature sensor</b>	Internal temperature

The failures dataset contains the precise times each gadget failed. During the course of the week, 105 failures were recorded. Device failure is to be classified when the time remaining to device failure is less than one hour.

For more information on the dataset and use case, please see [] <https://github.com/Unikie/predictive-maintenance-tutorial>

## 4.2 Algorithms

### 4.2.1 Support Vector Machines

Support Vector Machines (SVM) is a common supervised machine learning algorithm for both classification and regression tasks. A key attribute of SVMs is its high accuracy and precision in the segregation of classes. SVMs create  $n$ -dimensional hyperplanes to segregate datapoints into  $n$  number of classes/groups. The algorithm aims to achieve the maximum margin between support vectors (closest points), i.e. maximise the minimum margin.

In the case where two classes can be linearly separated, we consider the following as a representation of a dataset,  $S$ .

$$S = \left\{ x_i \in \mathbb{R}^{1 \times p}, y_i \in \{-1, 1\} \right\}_{i=1}^n \quad (4.1)$$

The values  $\{-1, 1\}$  represent two classes of data,  $A$  and  $B$ ,

$$y_i = \begin{cases} 1, & \text{if } i\text{-th sample} \in A \\ -1, & \text{if } i\text{-th sample} \in B. \end{cases} \quad (4.2)$$

The hyperplane can then be defined as  $F_0$  in  $\mathbb{R}^D$  space as,

$$F_0 = \{x | f(x) = x\beta + \beta_0 = 0\} \quad (4.3)$$

where,  $\beta \in \mathbb{R}^D$  with norm  $\|\beta\| = 1$

For a new sample  $x^{new}$  which is not within dataset  $S$ ,

$$y_{new} = \begin{cases} 1(\text{Class A}), & \text{if } f(x^{new}) > 0 \\ -1(\text{Class B}), & \text{if } f(x^{new}) < 0 \end{cases} \quad (4.4)$$

### 4.2.2 k-Nearest Neighbours

k-Nearest Neighbours (k-NN) is a simple supervised machine learning algorithm used in both classification and regression problems. This approach classifies objects based on the computational distances or similarities between samples/values. The k-NN algorithm only requires tuning of a single parameter,  $k$ , which represents the amount of nearest samples within the neighbourhood. The choice of  $k$  will affect

the algorithm's performance where a value too small would create higher variance hence resulting in less stability. A larger  $k$  value will produce higher bias resulting in lower precision.

After the number of neighbours,  $k$ , has been selected, the distances between the query data point,  $x_q$ , and an arbitrary data point,  $x_i$  are to be determined. Most commonly used is the Euclidean distance (4.5), however Manhattan distance (4.6) may also be applied.

$$d(x_q, x_i) = \sqrt{\sum_{i=1}^m (x_q - x_i)^2} \quad (4.5)$$

$$d(x_q, x_i) = \sum_{i=1}^m |x_q - x_i| \quad (4.6)$$

The resulting values are then sorted by distance from smallest to largest and the first  $k$  entries are selected. In classification problems, the mode of  $k$  labels is returned, while the mean of  $k$  labels is returned in regression problems.

### 4.2.3 Random Forest

### 4.2.4 Neural Networks

## 4.3 Results

## 4.4 Recommendations

Literature review:

Regression application - Prediction of Remaining Useful Lifetime (RUL) of Turbofan Engine using Machine Learning

# Reward Hacking



# Conclusion

# References

- [1] Q. Chen, W. Wang, F. Wu, S. De, R. Wang, B. Zhang, and X. Huang, “A survey on an emerging area: Deep learning for smart city data,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 5, pp. 4392–4410, 2019.
- [2] B. Mohapatra, “Machine learning applications to smart city,” *ACCENTS Transactions on Image Processing and Computer Vision*, vol. 5, pp. 1–6, 02 2019.
- [3] D. Luckey, H. Fritz, D. Legatiuk, K. Dragos, and K. Smarsly, “Artificial intelligence techniques for smart city applications,” 08 2020.
- [4] I. Saif and B. Ammanath, “trustworthy ai is a framework to help manage unique risk..” MIT Technology Review, 2020.
- [5] R. Moutafis, “How bad facial recognition software gets black people arrested..” towardsdatascience, 2020.
- [6] B. E, L. R. Flaih, D. Yuvaraj, S. K, A. Jayanthiladevi, and T. S. Kumar, “Use case of artificial intelligence in machine learning manufacturing 4.0,” in *2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, pp. 656–659, 2019.
- [7] S. Saria and A. Subbaswamy, “Tutorial: Safe and reliable machine learning.” ACM Conference on Fairness, Accountability, and Transparency (FAT\* 2019)., 2019.
- [8] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Man, “Concrete problems in ai safety,” Jul 25 2016.
- [9] H. Jiang and O. Nachum, “Identifying and correcting label bias in machine learning,” Jan 15 2019.

- [10] D. Jedamski, “Bias/variance tradeoff - applied machine learning: Foundations. linkedin learning..” Available at <https://www.linkedin.com/learning/appliedmachine-learning-foundations/bias-variance-tradeoff?u=2129308>, 2019.