Deanonymising Households Trading on a Blockchain Smart Grid

ANDREW MATHER

Supervised by Professor Raja Jurdak and Dr Ali Dorri.
Science and Engineering Faculty, Queensland University of Technology, 2 George St. Brisbane City, QLD, 4000, Australia.

September 6, 2020

This project aims to contribute to the study of user anonymity in blockchain for the Internet of Things. The research will explore machine learning methods to deanonymise users on a smart grid with blockchain. Data stored on decentralised blockchains is permanent and user privacy can be at significant risk. Research has found anonymity concerns in blockchain but IoT and smart grid contexts warrant further research.

The project analysis will include stored blockchain transactions and off-chain weather data. Section 1 will investigate the research topic and the surrounding literature. Followed by covering the methodology to complete the project. The aim is to determine how effective machine learning is to deanonymise smart grid blockchain users. This will highlight long-term anonymity risks of using blockchains due to permanent transaction data.

The approach will first, source energy grid data and construct appropriate blockchain ledgers. Second, apply machine learning analysis on these blockchains to identify households and their locations from past data. Third, incorporate off-chain solar exposure data as households also produce solar energy. Last, the project will investigate and test obfuscation methods to improve user privacy. This will test varying public keys and timestamps on transactions. Project outcomes include a better understanding of user privacy risks in a blockchain smart grid scenario but also methods to mitigate these risks.

The progress reported so far is the completion of creating blockchain ledgers, applying transaction classification, and adding off-chain solar exposure data. There were four classifiers tested and the best accuracy of about 79% was achieved with a convolutional neural network on hourly transaction data. This occurs when user's take limited steps to protect their privacy. Adding solar data showed increases for all tests but the best performance increased accuracy to 82%. Experiments with different security scenarios user's can implement is the final part to complete.

CONTENTS

1. INTRODUCTION

1.1. Introduction

The massive growth in the Internet of Things (IoT) to collect, process, and send data via the Internet, requires a framework to handle all these devices. The IoT plays a role in many applications, for example, 'smart' devices in households, energy grids, and smart cities. Centralised IoT systems are challenged by cost, efficiency, and security as their size grows. A decentralised approach to the IoT's considerable mass of information [1] will become required, but user privacy and security challenges should be addressed.

Blockchain can handle this data as a decentralised ledger to record transactions carried out in a network. This is a developing field with wide potential uses. Applications include cryptocurrency, financial systems, smart contracts, and non-monetary areas such as IoT and smart grids. Blockchain creates a level of anonymity for users through cryptographic means using private and public keys (PK).

The level of user anonymity from a permanent ledger is not studied in-depth in an IoT setting, despite the growing use of blockchain it. Studies on blockchain reveal malicious nodes can compromise user anonymity by classifying transactions using machine learning (ML) [2, 3] and off-chain data [3]. The project aims to contribute to the study of user anonymity in an smart grid using blockchain. It will explore methods to deanonymise users using ML to analyse transactions [4] and incorporate off-chain solar data [5].

1.2. Thesis Statement

The project will investigate user anonymity in smart grid blockchain transactions by using machine learning to classify households and their location. Classification will be aided by incorporating off-chain solar exposure data which can be compared to household energy production. The purpose is to link transactions to users and identify their 'ID' and location to deanonymise them. Techniques to enhance a user's privacy will also be suggested and measured.

1.3. Context and Aim

Studies on blockchain reveal malicious nodes can compromise user anonymity. Such a node can link similar transactions and also use off-chain data, such as, weather data. Research has not studied user anonymity in IoT blockchain in-depth, despite widespread use. There is complexity introduced by time series data and linking transactions as unique public key use increases.

The project aims to determine how effective ML is in deanonymising users in a smart grid implementing blockchain. This should highlight long-term anonymity risks of blockchain by analysing permanent transaction and historic weather data. The research is limited to a smart grid setting and ML classification as the analysis method. It is also important to determine and measure methods to improve user anonymity.

1.4. Objectives

Objective 1: Populate blockchain ledgers from energy data.

- Source appropriate energy grid data.
- Convert to a blockchain format suitable for analysis.

Objective 2: Find the success rate of classifying blockchain transactions as specific users or locations.

- Measure the likelihood to link and classify a user's set of transactions.
- Investigate what classification models are effective.

Objective 3: Find the success rate of when off-chain solar data is included.

- Measure the likelihood to link and classify a user's set of transactions.
- Measure the likelihood a user's energy production correlates directly to solar data.

Objective 4: Investigate the effectiveness of techniques to improve user anonymity.

- Investigate methods to increase user privacy.
- Measure the effect of obfuscation techniques varying transaction public keys and timestamps.

1.5. Significance

Blockchain for IoT has attracted tremendous attention recently. A huge volume of personalised data will become permanently stored in blockchains. Thus, it is critical to study the anonymity of the users in IoT. Identifying users and linking them to transactions in a smart grid, not only reveals private information, but also information such as when a home is unoccupied.

This research is undertaken to achieve:

- A better understanding of risks in adopting blockchain for smart grids.
- Objective 2 will establish the likelihood an attacker can link and extract a user's blockchain data.
- Objective 3 will establish the likelihood an attacker's success is increased using weather data.
- Objective 4 will analyse a range of privacy improving methods and measure their effectiveness.

The research is undertaken from an attacker's perspective on smart grid using blockchain. The project involves trying to deanonymise households in the blockchain which a standard user would not attempt. The outcomes aim to benefit future users and ensure their privacy in an emerging technology.

2. BACKGROUND AND LITERATURE REVIEW

Literature from key areas of the project will be reviewed to highlight key concepts, locate information to aid the project's completion, and identify a research gap in section 2.6. First covered is background information regarding blockchain and it's role in IoT and smart grids. Next, similar prior works in user anonymity and privacy are discussed. Last, relevant research is presented on machine learning classification techniques, in particular, for time series data.

2.1. Blockchain

Blockchain is a framework to create a public and universal distributed ledger. Bitcoin introduced blockchain [6] as a transaction ledger to ensure auditability, immutability, and non-repudiation. Blockchain implements a method to reach consensus between unreliable parties. Whereas a standard process has a trusted third party (TTP), like a bank, responsible for transaction security. Blockchain properties remove the need for TTPs. Blockchains store ordered transactions in blocks that are linked to a previous block. Blocks contain a header, with a unique ID, and information [1]. Each block header stores the preceding block's hash to establish the links.

Network participants managing a blockchain are nodes or miners. They collate transactions into blocks to append to the blockchain. Networks use consensus algorithms to maintain trust and agreement to add a block. For example, Bitcoin uses a Proof of Work algorithm [6], and Ethereum Proof of Stake [7]. Transactions use encryption, hashes, and public key (PK) cryptography. Digital signatures encrypt a document hash, signed with private keys, and PKs prove who signed it [1]. Blockchain participants create anonymity with PKs concealing their identity. Changing PKs between transactions, as in Bitcoin [6], can improve user anonymity.

2.2. Blockchain for the Internet of Things

The IoT is made of physical devices connected to the internet which use communication networks to process data [1]. It makes devices 'smart' and gain computation and communication capabilities. Many devices cause large data traffic [8] and future applications could reach billions of devices [1]. Challenges for IoT devices include limited computing power, storage and bandwidth, and data bottlenecks. Blockchain can change how IoT networks operate with a decentralised framework [9, 10]. IoT networks could enjoy blockchain's lower costs, decentralised management, and inherent privacy [8]. A combination of IoT and blockchain looks to solve the challenges faced in IoT networks [10]. There are many IoT applications in daily life, businesses, and society, shown in Figure 1. Intelligent power distribution, or smart grids are relevant to this project.

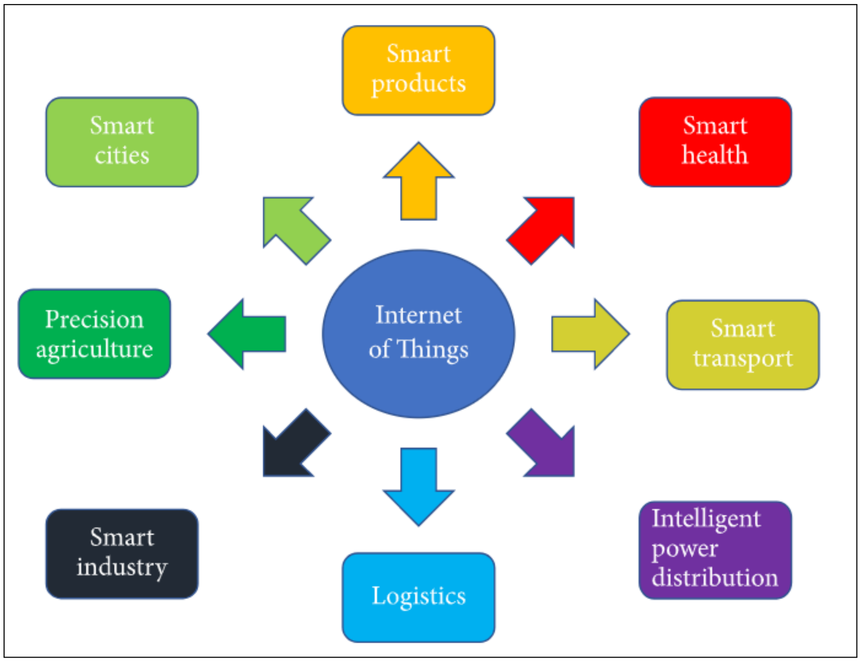


Fig. 1. IoT applications [1]

Energy systems and trading are developing quickly with the benefits offered by the IoT [11]. Smart grids allow systems to communicate and optimise energy production, consumption [12], and thus utilisation [13]. They create the infrastructure to transfer energy between distributed producers and consumers. Some consumers can also be energy producers (prosumers) with solar energy [11]. Continuously growing energy demand and supply has led to a desire for a decentralised energy system [14]. The authors at [11] highlight the challenges for such a system. First, managing transactions between users and the grid. Second, fluctuating supply from distributed renewable energy sources (e.g. rooftop solar). Third, TTPs lead to less efficient energy systems with more errors and operation costs. Blockchain is a likely path forwards to solve these issues in a smart grid [15].

Smart grids can decentralise behaviour using a blockchain framework. Blockchain and consensus algorithms can automate transactions and avoid TTPs. Other benefits can extend to real-time trading, anonymity features, and lower costs [16, 17]. The limits of IoT devices earlier are obstacles in implementing this system. The authors in [18] proposed a decentralised energy supply architecture which provided on-demand energy for miners in an IoT network using microgrids.

2.3. Anonymity Concerns

Using the IoT has many benefits but also increases exposure to new types of security and privacy threats. IoT issues go beyond standard information and privacy concerns as it can relate to people's physical lives and security. Privacy relates to the large amounts of personal data used by smart devices [1]. Likewise, smart grids will be complex networks, leading to privacy concerns [19] that need new approaches to solve [20].

Blockchains can solve problems of centralised systems and increase resilience to failures and attacks [19]. This does not preclude new types of privacy risks. Blockchain anonymity research has been primarily into digital currencies [21, 22]. With its attractive properties, however, to create a trusted smart grid [19] further research is required in non-monetary IoT anonymity.

Blockchain users create auditability through PKs while maintaining anonymity. The purpose is to mask a user's transactions, purchases, or information [1]. Relevant to a smart grid would be a participant's energy purchasing amounts, and times. When a household is not consuming energy implies the house is unoccupied. Privacy in blockchain should maintain transaction anonymity and have no ability to untie transactions [1]. Transaction anonymity means a transaction cannot be linked to a user, this is where different PKs are relevant. Untying transactions means transactions are not bound to user identities after routed through the network.

2.4. Related Works - Extended

We can apply ML approaches to a blockchain to investigate if user transactions are linkable. As noted earlier, blockchain users have PKs on transactions to achieve anonymity, with more public keys improving this. With supervised ML as suggested by [2], a malicious node could deanonymise a user by classifying transactions. An attacker can use the flow of inputs and outputs, to link user transactions. An attacker can attempt deanonymisation with real-time network traffic, but this research will focus on blockchain and historic weather data. IoT networks are subject to privacy risks around the exposure of user activity patterns from sensed data [2].

The authors of [2] concluded cryptocurrency studies show users can be deanonymised from transaction patterns stored on a blockchain. Their research analysed blockchain transactions in an IoT and smart home environment with ML to classify devices. Analysis was performed as an informed and blind attacker on devices within the smart home. The attacker's aim was to link transactions to their type of smart device. For example, identifying which transactions belong to a smart lock, thus inferring when an owner leaves their home. They populated a blockchain from real-world smart home network traffic. Then the attack method monitored the frequency of device transactions, using ML algorithms to compare with known frequency patterns of potential devices. Results showed an informed attacker being up to 90% accurate, and a blind attacker around 30%. This indicates a serious risk in the privacy of devices using the blockchain. [2] also proposed methods to improve user privacy in the IoT blockchain. Three timestamp obfuscation methods reduced successful device classification by up to 30%. The techniques were combining multiple packets into one transaction, merging ledgers, and adding random transactions delays. We can draw parallels between [2] and this project where smart devices become smart homes and we investigate the pattern of energy over time, as opposed to frequency.

Authors at [22] suggest an attacker in a multiple PK scenario needs to create a one-to-many mapping between users and addresses. The analysis process suggested by [22] involves three stages. First, the flow of blockchain transactions (nodes) in a transaction graph where PKs are the inputs and outputs. Second, the flow of payments between PKs in an address graph created from the transaction graph. Last, a user graph with the users and each PK that may belong to the same user; drawn from the previous information and blockchain heuristics. [23] used a full blockchain analysis to link users to public addresses.

For ML with unsupervised approaches, [3] is an example of clustering blockchain addresses. Clustering is a valuable method for ML problems [24], related to splitting data into groups. [3] takes a clustering approach to blockchain transactions and also off-chain data. Their scenario showed successful clustering of information, with off-chain data improving the accuracy. Clustering household energy patterns may show similarities between households located nearby.

[25] developed a method to deanonymise blockchain transactions using supervised machine learning to predict new entities. They perform multi-class classification to categorise a cluster of transactions. They use decision trees with random forest and gradient boosting algorithms. This paper has parallels with the project

in attempting to categorise sets of transactions as belonging to different households in a smart grid.

The authors of [26] desired an accurate approach to compare weather data and power generation. They introduced linear and nonlinear time models for solar intensity prediction. This method could be a useful technique for the project when comparing solar data to a user's energy production. Additionally, [27] explores techniques to convert daily solar data into hourly information by modelling the pattern over a standard day. This is important as the project's solar data was only available at daily resolution.

A work relevant for the obfuscation techniques part of the project is [28]. The authors propose a privacy-preserving and data aggregation scheme. This will be useful for potential obfuscation techniques as they discuss dividing users into separate blockchains (ledgers) and using multiple pseudonyms (public keys) to protect a user's identity. There are similarities in the nature of these methods with [2].

2.5. Time Series Classification

Classification problems with data that can be ordered, can be treated as a time series classification (TSC) problem [29]. Researchers have investigated many methods to effectively classify time series data [30]. Popular are nearest neighbour classifiers with a distance function if appropriate [31]. [31] also shows an ensemble of classifier's outperforms the individual components. These approaches use either an ensemble of decision trees (random forest) [32] or an ensemble of different types of discriminant classifiers [33].

[29] gives an overview of potential deep learning applications for TSC. The authors found for univariate and multivariate data, the top three types of networks were residual (ResNet), fully convolutional (FCN), and mulitlayer perceptron (MLP) networks. The MLP is traditional form of deep neural networks and was proposed in [34] as a baseline architecture for TSC.

Random forest is a decision tree machine learning approach used by [35] and [25] for TSC. The authors of [35] compared the effectiveness of different decision tree approaches for TSC. They found support vector machines performed poorly and ensembles are favoured when after optimal accuracy. The better performing ensembles were MultiBoost and AdaBoost.M1. However random forest performed similarly well and would be more favoured on larger datasets. This project will look to use a decision tree for classifying user blockchain transactions and will consider these options.

Correlation and cointegration are potential statistical approaches to compare time series. This is relevant for the project's comparison of household energy transactions to solar data. The authors in [36] investigated how to optimise wireless sensor networks in environmental monitoring. They used a statistical approach to cointegrate sets of time series data to select the optimal number of sensors. This was successful showing only 25% of the original sensors were not cointegrated. In particular [36] describes an analytical framework to analyse multivariate time series data which relates to this project comparing solar data to user transactions.

2.6. Research Gap Identified

Research in blockchain user anonymity is developing and uses both transaction and off-chain analysis. Despite widespread usage of blockchain in IoT, user anonymity level is not yet studied thoroughly. The literature investigated shows research into IoT implementations of blockchain and some into the privacy, usually Bitcoin focused. The combination of machine learning analysis on stored data in the smart grid context is a new contribution. Privacy concerns abound as smart grids develop and it is important to understand the risks before storing user data on a permanent and public ledger.

3. METHODS AND PLAN

3.1. Research Process - Updated

The following contains the research process broken down into steps. It covers the tasks required to address the thesis statement and objectives. Figure 2 shows an overview of the main phases planned for the project.

1. Background information and literature review on relevant papers for:
   (a) Blockchain-based IoT, smart grid and energy trading.
   (b) Privacy concerns in IoT blockchain contexts.
   (c) Machine learning classification techniques.
2. Source an appropriate energy dataset. It should contain a reasonable number of customers, energy use, solar energy production, and information that distinguishes users in different locations.
3. Convert energy data into blockchain ledgers for objective one. Include the ability to adjust the frequency of transactions, number of ledgers, and number of public keys per customer.
4. Perform machine learning analysis on blockchain transaction data for objective two.
   (a) Test a variety of classification models.
   (b) Apply to classifying transactions by customer or location.
   (c) Measure with respect to transaction frequency.
5. Perform machine learning on blockchain with off-chain solar exposure data for objective three.
   (a) Combine user datasets with historic solar exposure data to increase attacker accuracy.
   (b) Perform include statistical comparisons between user energy production and solar data directly.
6. Suggest and evaluate methods to improve user anonymity for objective four.
   (a) Measure the change in privacy as consumers use additional public keys.
   (b) Measure the change in privacy as ledgers are mixed.
   (c) Measure the effectiveness of timestamp obfuscation.
7. Deliver progress and final project reports.

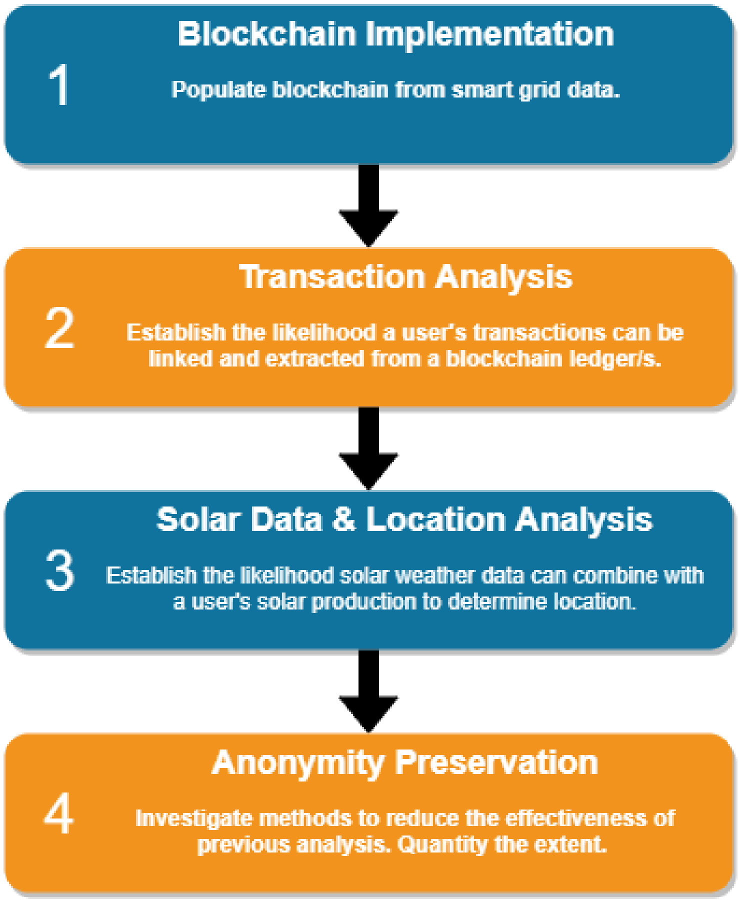

1
Blockchain Implementation
Populate blockchain from smart grid data.
2
Transaction Analysis
Establish the likelihood a user's transactions can be linked and extracted from a blockchain ledger/s.
3
Solar Data & Location Analysis
Establish the likelihood solar weather data can combine with a user's solar production to determine location.
4
Anonymity Preservation
Investigate methods to reduce the effectiveness of previous analysis. Quantity the extent.

Fig. 2. Phase design

3.2. Technical Frameworks

- Analysis techniques:
  - Classification with decision trees and neural networks.
  - Correlation and cointegration.
- Python (3.8):
  - Data manipulation: Pandas (1.1.0) and NumPy (1.18.5).
  - ML: Scikit-learn (0.23.1) and Keras (2.4.3) with TensorFlow back end (2.3.0).
  - Graphing: Matplotlib (3.3.0) and Seaborn 0.10.1.

3.3. Data Collection - Updated

Past energy use and generation data is sourced from Ausgrid solar home electricity data [4]. It contains Australian household data where households are prosumers. The data set has half-hour data from 1 July 2010 until 30 June 2013 for 300 households across New South Wales. It includes energy consumption (on and off-peak) and generation. All households have a full data set and Ausgrid performed quality checking. Other datasets found such as [37] include less households making anonymity hard to study, and are not situated in Australia with available solar data from the Bureau of Meteorology.

A blockchain will be populated with transactions corresponding to the energy use and generation of households in the dataset. Different blockchains will be made with transaction frequencies of per week, day, hour, and half-hour. The highest frequency is half-hour periods the data provides. Each energy use period will be treated as a communication between a smart meter and the grid. Thus, the process will generate a transaction for each period per household. Assumptions for this process will include:

- Real-time network traffic will be abstracted out. The focus is on attackers with access to permanently stored transaction information.
- Blockchain algorithms such as consensus algorithm are not required and will assume to pass each transaction. Patterns of transactions are not reliant upon these blockchain steps.

A single node will act as a miner collecting all transactions until a blocksize is reached. Then the miner creates a new block appended to the ledger. The results will be analysed with ML algorithms to deanonymise households to achieve objective two. Relevant off-chain data is required for the second research objective. Ausgrid data provides postcodes for each household and therefore historical data from the Bureau of Meteorology at [5] is easy to source and is accurate.

3.4. Data Analysis

Each phase of the project will perform a similar ML analysis on the acquired data. The project will use Python frameworks to perform mainly classification. Python and the required libraries listed in 3.2 are freely available and appropriate.

The method will have an 'attacker' training machine learning models locally and measuring the ability of these models to predict users and their location. The next stage will use further classification and statistical approaches to link a user to the most similar solar data set. The last stage investigates obfuscation methods to increase resilience against ML attacks used. Likely suitable will be varying public key numbers, mixing ledger, and timestamp obfuscation methods.

3.5. Project Management

To prepare to complete this project, the key research questions, objectives, and outcomes were detailed to highlight the focus of the project. A literature review was completed to reinforce the project direction and goals. This was extended to develop reasoning behind the selected classification methods implemented. A break-down of the phases of the project defines the necessary order of work in Figure 2 earlier. These processes will aid to ensure the project is completed in manageable sections and with appropriate quality control and testing. Each step shows the main research and investigation activities required.

3.6. Project Timeline - Updated

Figure 3 provides an overview of the project timeline and progress.

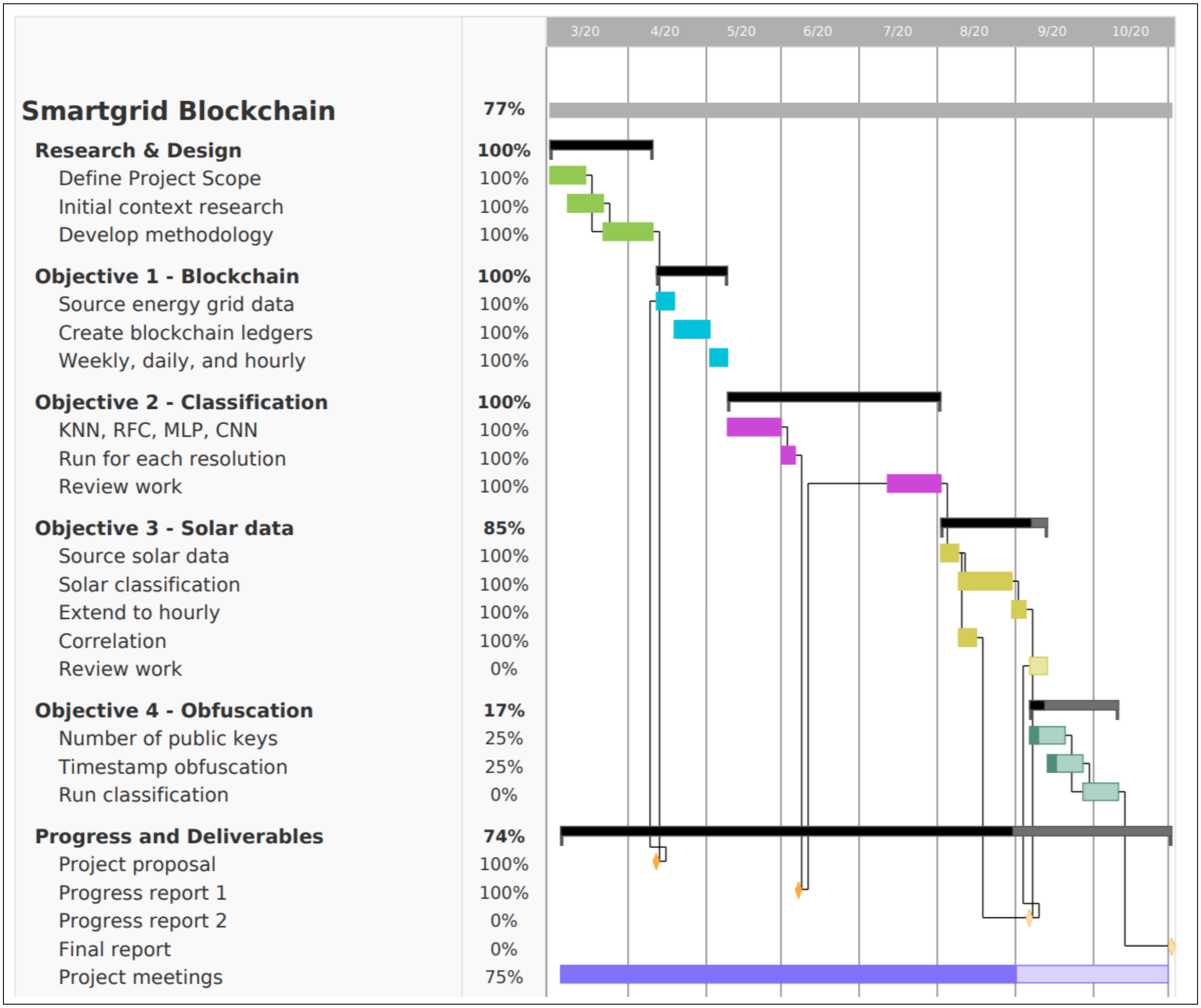


Fig. 3. Project timeline and progress

4. RESEARCH RESULTS

4.1. Objective One - Populate Energy Grid Blockchain - Revised

Energy grid data

The first step after project design and proposal was to source appropriate energy data. Ausgrid energy data [4] was found and evaluated to suit the project. It contains all the information desired to complete the research. The following Table 1 shows a sample of the dataset.

Table 1. Example original energy data

Customer	Generator	Postcode	Type	Date	0:00	0:30	1:00	...	23:30
1	3.78	2076	CL	01/07/2013	1.250	1.244	1.256	...	1.081
1	3.78	2076	GC	01/07/2013	0.303	0.471	0.083	...	0.068
1	3.78	2076	GG	01/07/2013	0	0	0	...	0

This dataset contains the energy use and solar production of 300 households over three years. The data frequency is in half-hour blocks which is nice from the perspective of creating blockchain transactions of a reasonable resolution. Each day a household's data is split into off-peak consumption (CL), general consumption (GC), and gross solar generation (GG). Additional features are the household generator size and postcode. Generator sizes range from 1kWh to 10kWh systems, but average low in this range at 1.68kWh. This is a key driver of energy production and useful to have available. Important are the user ID and postcode attributes allowing classification analysis from data to predict user or location (aided by off-chain data). Solar exposure data by area, in the time period of the energy data is available at [5] and discussed in section 4.3.

Wrangle energy data

Sourced energy data needs to first, be wrangled into a suitable format for time series classification, and second, have the required features of blockchain ledgers added. This section describes the process to populate a blockchain from the energy data with various options. Visualisations of the dataset are provided and used to explain how a household can be 'fingerprinted'. An attacker is able to use transaction timestamps, values, and PK information to link data for each and possibly between PKs in the blockchain.

The data was manipulated into a time series format with a date-time column removing the separate attributes for each time period. This allows each blockchain ledger row to represent one transaction. Blockchains were produced with different transaction time frequencies; half-hourly, hourly, daily, and weekly. Larger transaction periods may more distinctly identify users, and work against overfitting, however, there is far less data to identify a user. At this stage, the three years were left separate and 0 amount values kept. Table 2 shows a sample of the rearranged data.

Table 2. Example wrangled energy data

Customer	Postcode	Type	Datetime	Amount
1	2076	CL	01/07/2013 0:00	1.250
1	2076	GC	01/07/2013 0:00	0.303
1	2076	GG	01/07/2013 0:00	0
1	2076	CL	01/07/2013 0:30	1.244
1	2076	GC	01/07/2013 0:30	0.471

Create blockchain ledgers

The energy data is now suitable for creating blockchain ledgers. There will be four transaction types:

- Genesis transaction → First transaction for a ledger or public key.
- On-peak consumption (CL).
- Off-peak consumption (GC).
- Solar energy export (GG).

Adding several features is required to populate the blockchain. Each transaction has a hash (of its content) included as an identifier, and the previous transaction's hash to create a chain. Also each household signs transactions they generate with a PK. This produces the following structure of a transaction:
Hash | Previous Hash | Public Key | Timestamp | Transaction Type | Amount

In the initial stages of the investigation, three classification ledger scenarios are considered:

- One ledger per customer (LPC).
- One ledger per postcode (LPP).
- All one mixed ledger (AOL) with unique public keys per transaction.

First, ledger per customer allocates each customer's transactions to a separate ledger. Second, ledger per postcode groups households in the same postcode to a ledger but are still differentiated by their PKs. These two scenarios will have one PK per customer, representing limited security measures for a user. Last, with one fully mixed ledger and unique PKs per transaction is a difficult case for an attacker. Neither of these are realistic but will produce a bound of expectations. Realistic and privacy increasing (compared to best case) variations of ledgers and public keys will be analysed under objective four, obfuscation techniques. Households using a single public key is possible, but has a drastic reduction in privacy and is unlikely. On the other hand, new public keys require genesis transactions which cost the PK holder, thus a new PK per transaction is unlikely.

Table 3 shows a sample format of a created blockchain ledger. Note, customer ID and postcode are removed from the classifier training and test sets.

Table 3. Example blockchain ledger

Hash	PHash	PK	Customer	Postcode	Type	Datetime	Amount
Genesis		PK_1	1	2076	CL	01/07/2013 0:00	1.250
a	Genesis	PK_1	1	2076	GC	01/07/2013 0:00	0.303
b	a	PK_1	1	2076	GG	01/07/2013 0:00	0
c	b	PK_1	1	2076	CL	01/07/2013 0:30	1.244
d	c	PK_1	1	2076	GC	01/07/2013 0:30	0.471
e	d	PK_1	1	2076	GG	01/07/2013 0:30	0

Data Visualisation

It is important to understand trends in the household energy patterns. Figure 4 shows the pattern of energy use and generation of four customers over an example day. Figure 4 brings two key insights. User consumption more clearly distinguishes users, and all customers have a similar solar generation trend (expected from day/night cycles) but magnitude depends on generator capacity. Customers shown, 37, 59, 102, 226 have solar capacities of 1.5, 2.8, 2.0, and 1.5kWh respectively. This explains why customer 59's production is greater.

Figure 5 shows the same household's consumption by day across the data. It is observed again consumption is a much better 'fingerprint' of a user's transactions. Trends, peaks, and troughs provide distinguishing features amongst these consumers. The generation is also more distinct in this view than the single day in Figure 4.

Figure 6 shows the consumption of the same consumers by week across the data. This view is quite similar to Figure 5 sharing many features, but with less fine variations and noise. While this visually looks to distinguish users easier, lost detail and reduced data is important, as section 4.2 will show.

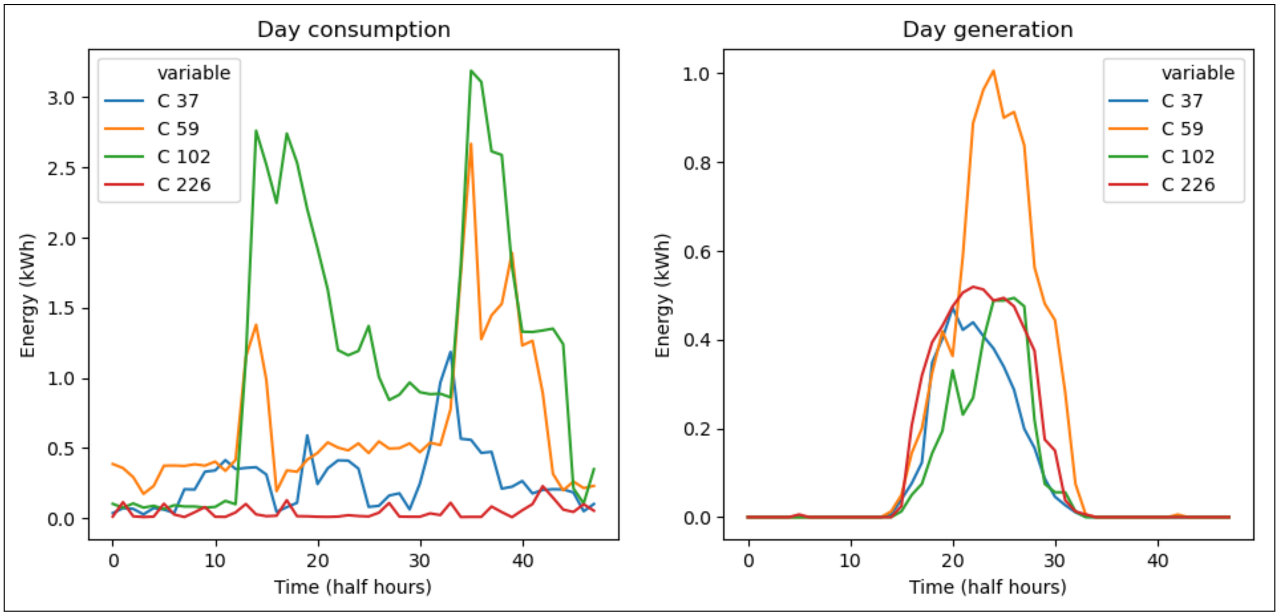


Fig. 4. Energy pattern of four customers over a random single day

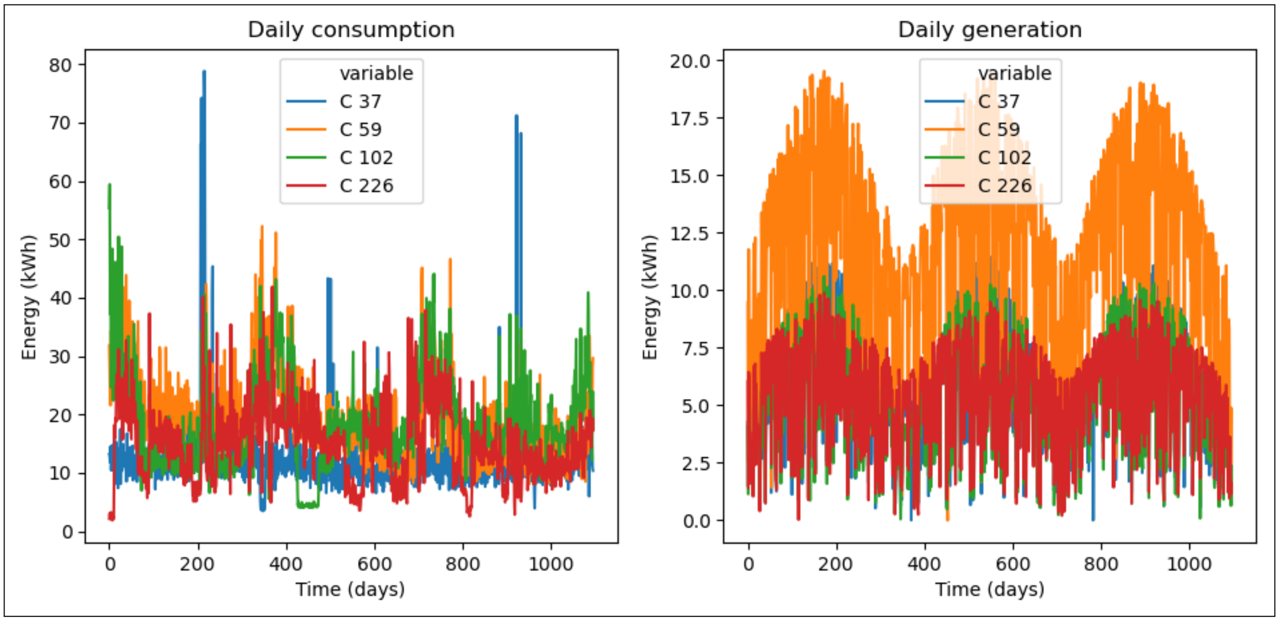


Fig. 5. Energy pattern of four customers by day over three years

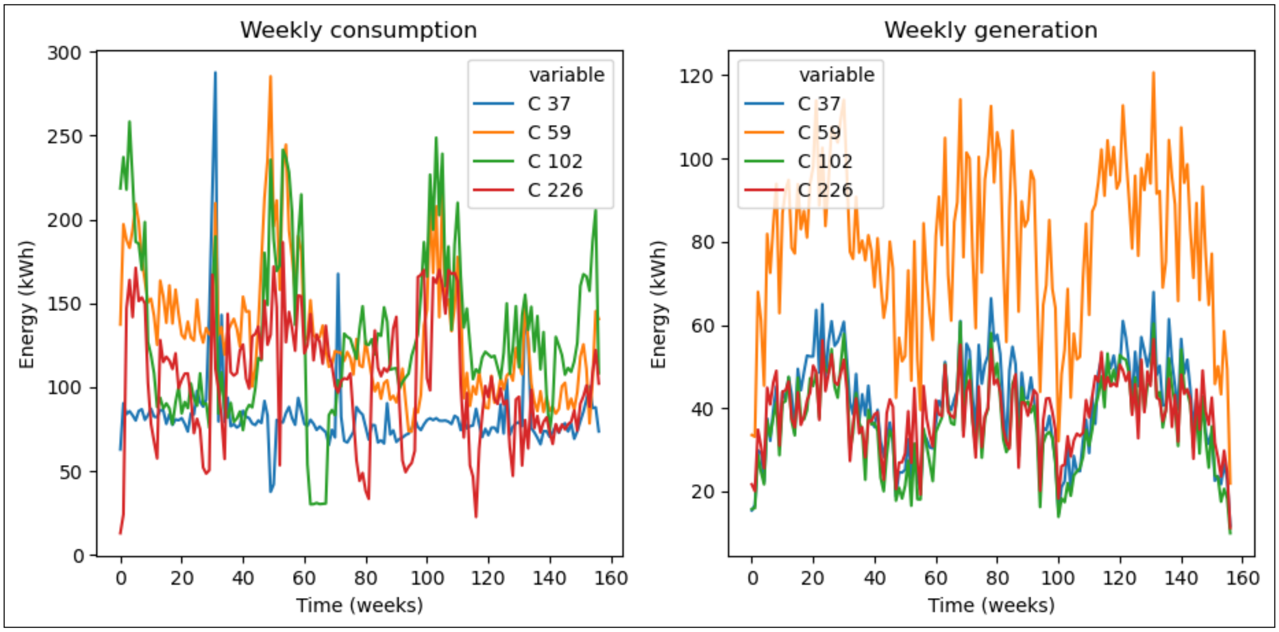


Fig. 6. Energy pattern of four customers by week over three years

4.2. Objective Two - Transaction Classification Methods - Revised

Attackers can construct a set of a user's energy transactions by linking transactions emerging from the same PK and those with statistical similarity. This provides a time series which can fingerprint a user's energy consumption and generation. The visualisations indicate consumption is likely more important than production to classify users in this stage. Solar generation, however, will matter more for objective three in section 4.3 when off-chain solar data is added. With a set of user transactions, an attacker may be able to first, continue to link transactions to a user as an ongoing privacy risk, and second, potentially reveal household location with off-chain solar data.

Selection of classifiers

The classification methods implemented include two decision trees and two neural networks:

- K-nearest neighbours (KNN).
- Random forest classifier (RFC).
- Multilayer perceptron neural network (MLP).
- Convolutional neural network (CNN).

The goal is to predict a category and the dataset has labelled data, thus classification is used over clustering. The dataset is not text based and a Stochastic Gradient Descent (SGD) or KNN classifier is suitable. A KNN classifier was selected as a simple baseline approach, initial testing easily outperforming a SGD model.

Section 2.5 discussed decision trees for time series classification (TSC). Random forest was an effective option, especially for large multivariate data like this project. Section 2.5 also covered suitable deep learning networks for TSC. CNN models were reported best in the main paper discussed, and MLP networks also suggested effective and may generalise larger multivariate data well.

Analysis approach

1. Preprocess data:

(a) Categorical data was made numerical and scaled.

(b) Random train and test sets (80/20) constructed. Customer, postcode, and generator dropped from train and test sets to leave only the blockchain data.

(c) Zero energy amount transactions removed as these would not create transactions.

(d) For the one ledger case PKs are made unique and ledgers combined.

2. Run each classifier for weekly, daily, hourly, and half-hourly transaction time frequencies to predict:

   (a) Customer: i) Ledger per customer (LPC), ii) Ledger per postcode (LPP), iii) All one ledger (AOL)

   (b) Postcode: i) Ledger per customer (LPC), ii) Ledger per postcode (LPP), iii) All one ledger (AOL)

3. Evaluate overall classifier accuracy. For the best model, top-5 accuracy will be measured.

4. Iterate model performance to tune hyperparameters for greater accuracy. For example, CNN filter size, number of trees in the RFC, or the ’k’ for KNN.

Transaction Classification Results

This section presents the results of the analysis outlined above. The best results were achieved by the CNN classifier, as suggested by literature. The results are presented in Figures 7, 8, and 9 separated by ledger scenarios previously listed. This allows easy comparisons of the classifiers and a discussion of each will follow. Appendix A contains graphs which separate the results by classifier and Appendix B has tabulated results.

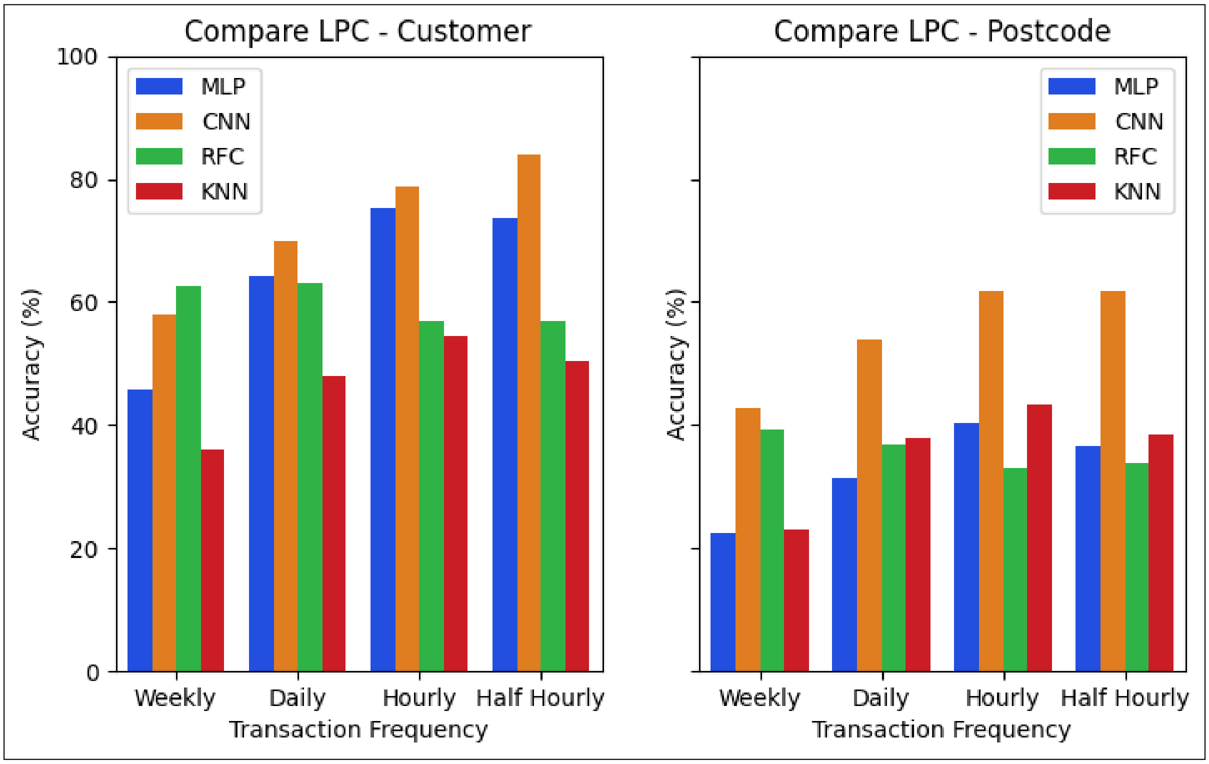


Fig. 7. Comparison of classification models on ledger per customer (LPC)

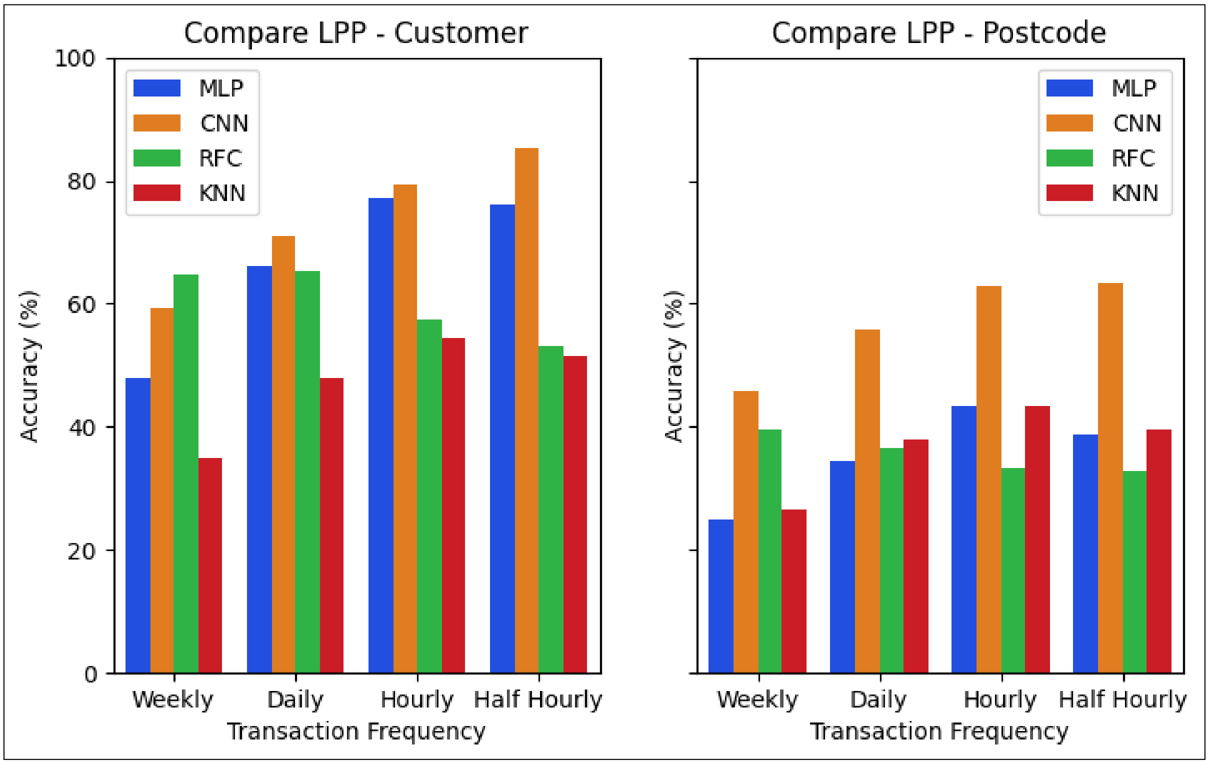


Fig. 8. Comparison of classification models on ledger per postcode (LPP)

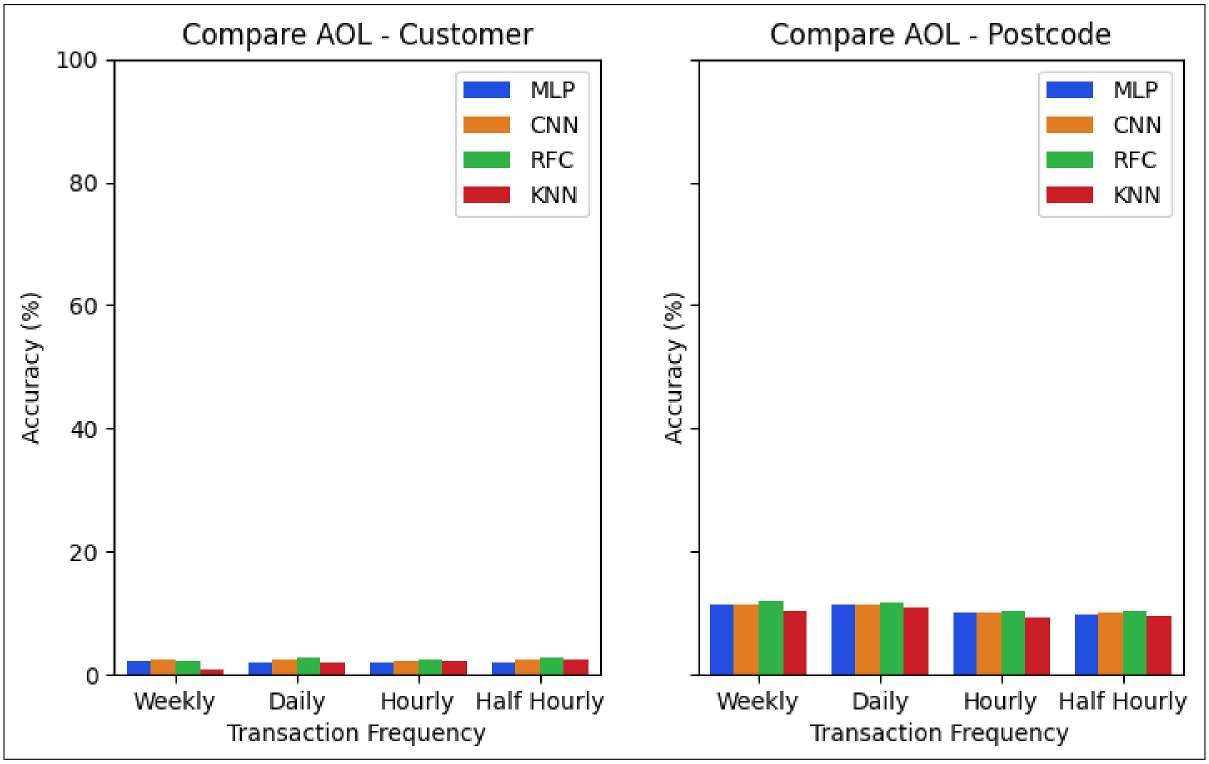


Fig. 9. Comparison of classification models on all one ledger (AOL)

K-Nearest Neighbours

A KNN was implemented first as a baseline classifier and had the least accurate predictions. The results show a maximum accuracy of 54% for customer and 43% for postcode classification on hourly transactions. This is an average and for example was 95% for predicting customer 138 but also 0% for a several users. These results are a promising starting point and were achieved after tuning the model 'k' value. Initial tests using $k = [1, 50]$ found the best results for customer classification with $k = 3$ and postcode classification $k = 2$.

There was no notable difference in performance between the LPC and LPP scenarios for the KNN model. In the mixed ledger case, customer and postcode predictions dropped to at best, about 2% and 11% respectively. A large decrease in accuracy is expected, however, the KNN model also performed worst here. An important trend in the results is the increase in accuracy as the time frequency increases to hourly, but drops at half-hourly. This smallest transaction frequency perhaps begins to mask consumer patterns, and overfitting occurs.

Random Forest Classifier

Next a RFC was implemented and the results are slightly better than the KNN as expected for an ensemble classifier. The best accuracies are 65% (customer) and 41% (postcode) on weekly data. The performance pattern is quite different as the model performs better with larger transaction frequencies. Weekly is the most accurate with it decreasing with each reduction in time period. Performance is better in the LPP scenario, which is reasonable as this ledger setup actually provides additional information. Users remain distinct by their PK but some grouping of users that share postcodes is known. The RFC mixed ledger case results are the best of all the models which may be important in later obfuscation tests.

An RFC classifier can provide the feature weights used by the model. The weights were similar in all tests and are 70% reliant on PK/ledger information, 20% on transaction amount, and 10% on transaction type. This seems reasonable as once a group of transactions is identified as a user, the remainder can be classified by sharing a PK. Changes in these weights will be interesting when PK numbers are varied.

Hyperparameters were tuned over test runs to establish better performance. The parameters considered were the number of trees, maximum tree depth, and the maximum features to consider when splitting. 100 trees were allowed to run till pure, unless memory limits were reached, with the square root of data attributes used for maximum feature splitting. The RFC results are a promising improvement as the neural networks are favoured by literature to perform best.

Multilayer Perceptron

The MLP classification results improve upon the RFC model significantly in the separate ledger scenarios. The results have a maximum accuracy of 77% (customer) and 43% (postcode) on hourly data. The LPP scenario slightly outperforms the LPC as similarly explained for the RFC model. Also the KNN pattern of accuracy loss at half-hourly data has occurred but to a lesser extent. The model struggles with the one ledger case and performs worse than both decision trees.

The results are averages and for example, 72 customers were classified with 100% accuracy, but also ten with 0%. This large range of outcomes is described by large standard deviations. For example for customer classification average standard deviations are 35%, 30%, and 10%, respective to LPP, LPC, and AOL scenarios. The spreads were slightly lower in more accurate transaction frequency tests.

The MLP model generally ran until convergence, but was limited to 1000 iterations. Hyperparameters tuned for the MLP model were the number of hidden layers and neurons per layer. Testing found best performance with three hidden layers of 10 neurons. Overall the MLP results are accurate and trend towards better accuracy with higher transaction frequency (until overfitting).

Convolutional Neural Network

The CNN classification model achieved the best results of all the methods by significant margin, except for the mixed ledger where the RFC outperformed it slightly. The results have a maximum accuracy of 85% (customer) and 62% (postcode) on half-hourly data. The CNN is the only model which improves with every increase in transaction frequency, including the half-hourly data. In particular, the CNN significantly outperforms other models in predicting postcodes, while only somewhat for customer outperforming the MLP. The one ledger case is handled similarly to the other models, perhaps signifying the neural network advantages, clear in the LPC and LPP scenarios, will lessen when more PKs and other obfuscation techniques are used.

Hyperparameters tuned for the CNN model were the filter size, batch size, and number of epochs. 128 was used for the filter and batch sizes, while 100 epochs were run for each test. Additionally, the CNN construction uses two 1D convolutional filters, and tested several optimisers and measures of loss for the best performance.

Top-5 accuracy tests were run for the CNN, as an attacker could rank several likely options to aid deanonymisation attempts. For the daily and hourly time frequencies the top-5 accuracy results are in table 4. This will be compared to a similar table after adding solar data in section 4.3. The results show almost perfect predictive power in the single PK (LPC and LPP) scenarios, and accuracy three-five times greater than overall accuracy for the mixed single ledger. These levels of accuracy show an attacker's chances at linking a user's transactions far above acceptable without obfuscation techniques. The top-5 accuracy is quite high even for the mixed ledger scenario and more sophisticated obfuscation may be required.

Table 4. **Top-5 CNN accuracy**

Frequency	Predictor	LPC	LPP	AOL
Daily	Customer	>99%	>99%	9.0%
	Postcode	97.7%	96.7%	30.2%
Hourly	Customer	>99%	>99%	9.4%
	Postcode	98.5%	>99%	30.5%

Overall Comparison

The CNN model performs the best on the data, especially for postcode prediction, and handles the more frequent data well. However, the RFC handles the low information single ledger scenario best and should continue to be considered. Most models trend towards better accuracy at more frequent transaction data, except can overfit at half-hourly. Throughout, LPP results outperform LPC as PKs still separate customers but additional postcode grouping information is provided. This will be interesting to see how it varies as PK numbers are changed.

Guesswork should expect average accuracies of 0.33% for customer and 1% for postcode (approximate as not evenly distributed). This is relevant for the low accuracy predictions in the AOL scenario. However, when the models perform well above this accuracy it does not follow predicting postcode should outperform user ID.

Overall machine learning models can quite accurately link user transactions from past blockchain data when users take limited steps to protect their privacy. More frequent transactions aid an attacker, as does using separated ledgers, whether by users or postcodes. The next section will highlight attackers can do even better by including off-chain solar data before considering user privacy enhancing measures.

4.3. Objective Three - Adding Off-Chain Solar Data - Added

Section 4.2 highlighted there is a high risk attackers can classify and link a user's energy transactions from statistical similarity. Users can take measures to reduce this likelihood, but first the possibility an attacker can add off-chain data to increase their success rate is considered. The physical nature of solar generation means, an attacker can source off-chain solar exposure (or other weather) data to aid their classification attempts. Solar exposure data can be split into areas over a smart grid and added as a feature to classification analysis, or statistically compared to solar generation transactions directly. The following will outline the analysis approach, sourcing and processing the solar data, and then discuss the results. The solar classification tests will use the CNN and RFC models established previously as performing well in different cases.

Analysis approach

1. Source solar exposure data for each postcode from the Bureau of Meteorology at [5]. This provided daily data available at a high frequency of locations.
2. Generate approximate hourly and half-hourly frequency data from the daily solar data.
3. Add time series solar exposure as a new feature to the dataset.
4. Run classification tests with CNN and RFC models:
    (a) Preprocess data as described by section 4.2's analysis approach.
    (b) Run each classifier for weekly, daily, hourly, and half-hourly transaction frequencies to predict:
        i. Customer: i) Ledger per customer, ii) Ledger per postcode, iii) All one ledger
        ii. Postcode: i) Ledger per customer, ii) Ledger per postcode, iii) All one ledger
    (c) Evaluate overall classifier accuracy using the same hyperparameters as section 4.2. For the CNN model, top-5 accuracy will be measured.
5. Statistically compare household data sets to each region of solar data. This helps show support for whether solar data should be beneficial to the previous analysis.
    (a) Correlate each household to all solar data sets.
    (b) Cointegrate each household to all solar data sets.

Solar data

A set of historic solar exposure data was sourced from the BOM at [5] for each postcode in the original energy data. Other weather data is also available but the analysis has been limited to the most relevant for solar energy production. The original data contains 100 unique postcodes across the 300 households and one set of off-chain data was collected for each. This process required determining the closest weather station to the centre of each postcode. All except three regional postcodes had a weather station within 5km but they are also larger areas far from others included. Several inner city postcodes do share the same closest weather station but this should have limited impact. All weather stations had solar data for the 2010-2013 time period required and Table 5 shows an example of the solar data format.

The solar data was only available in a daily format as more frequent data is only available at limited set of locations. Therefore a polynomial estimation approach discussed in [27] mentioned in section 2.4 was implemented. This uses a day's total solar exposure from the datasets and splits it into hourly pieces. This is being extended to half-hourly data but the results are not ready for this progress report. The daily and weekly analysis can directly use the data sourced.

Table 5. Example solar data

Year	Month	Day	Daily global solar exposure (MJ/m*m)
2010	7	1	9.9
2010	7	2	4.4
2010	7	3	10.6

Solar Classification Results

This section presents the results of analysis with added solar data outlined above. The best results achieved were in line with section 4.2. The CNN model performs best in the LPC and LPP cases, with solar data slightly improving accuracy as expected. While the RFC model again achieved the best results in the AOL scenario also improved by the solar data. The results are presented in Figures 10, 11, and 12 separated by ledger scenarios. This allows comparison of the classifiers and a discussion of each will follow. Appendix B contains tabulated results of the presented graphs.

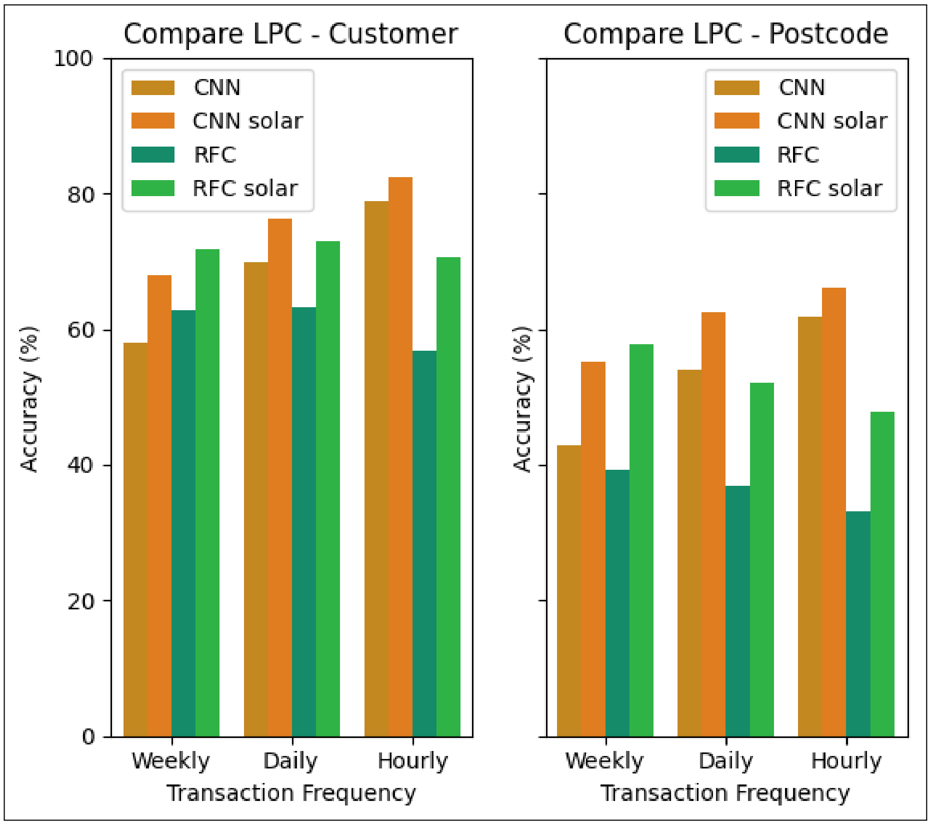


Fig. 10. Comparison of solar classification models on ledger per customer (LPC)

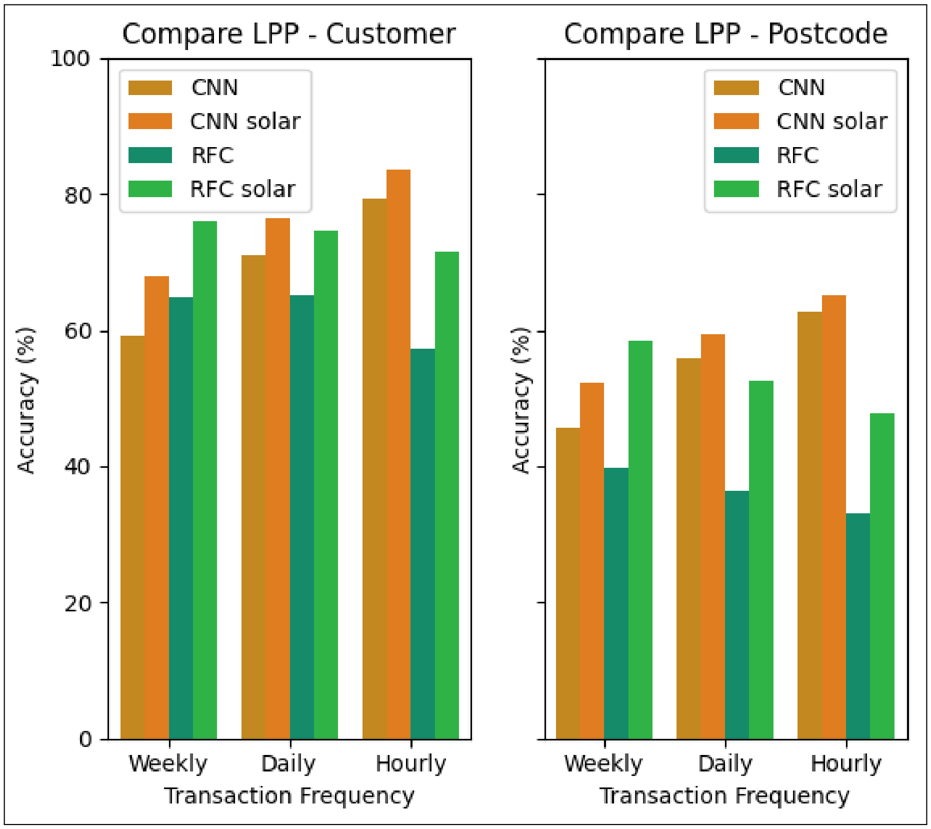


Fig. 11. Comparison of solar classification models on ledger per postcode (LPP)

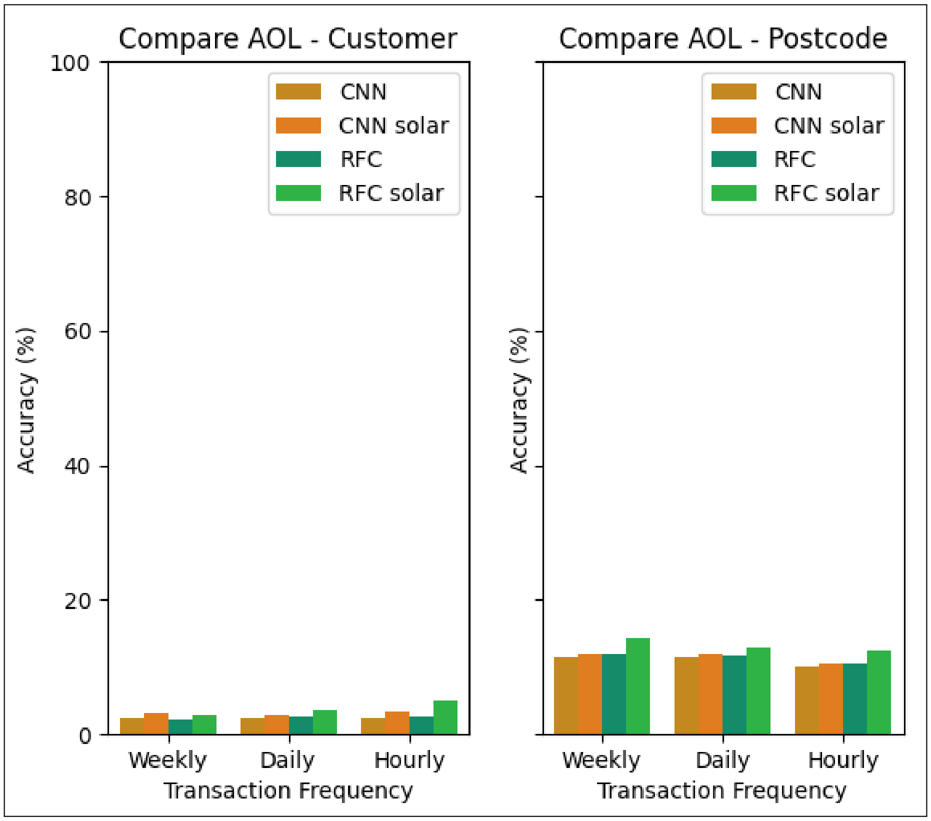


Fig. 12. Comparison of solar classification models on all one ledger (AOL)

Random Forest Classifier

The RFC was run first and the best accuracies are 76% (customer) and 58% (postcode). The results show an improvement in all cases and frequencies. The downward trend as frequency increases for LPC and LPP remains. The improvement on postcode prediction is greater than for customer in all scenarios. This would be expected as the solar data supports location prediction better. The model placed a 5% weighting on the solar data and the same parameters as section 4.2 were used.

Convolutional Neural Network

The CNN accuracy has also increased in all cases and frequencies with the inclusion of solar data. The best results of the tests run are 84% (customer) and 66% (postcode). The increasing trend as frequency increases for LPC and LPP remains. While results for half-hourly data are not yet available, it seems likely they will also improve and as in section 4.2 and have the greatest accuracy. The improvement on customer and postcode predictions are even, unlike the RFC which benefit far more in postcode prediction. This will be further considered as postcode prediction was expected to benefit most. The one ledger case is handled similarly to the other model, perhaps signifying the neural network advantages, clear in the LPC and LPP scenarios, will lessen when more PKs and other obfuscation techniques are used. Table 6 shows the top-5 accuracy predictions and the deltas to the equivalent table in section 4.2. Improvements here can only be seen in the one ledger scenario and these were small.

Table 6. **Top-5 CNN daily data with solar accuracy**

Frequency	Predictor	LPC	LPP	AOL
Daily	Customer	>99% (0%)	>99% (0%)	11.0% (+2.0%)
	Postcode	97.7% (0%)	>97.6% (+0.9%)	31.7% (+1.2%)
Hourly	Customer	>99% (0%)	>99% (0%)	11.5% (+2.1%)
	Postcode	>98.7% (+0.2%)	>99% (0%)	30.5% (0%)

The energy data set provides gross generation measured by the solar meter. However, in a blockchain energy trading environment, sale transactions would contain energy exported. Energy exported would be gross generation less household use of their production. This is a limitation of the dataset for the purposes of the analysis performed. Classification results would likely benefit from energy export (over generation) as it would better uniquely identify users and avoid the daily generation for same system sizes being highly similar. However, benefits from adding solar data may be lessened, especially in the following statistical comparison section, as the movements of energy export are affected by household use.

Solar Statistical Comparison

To support the solar data providing the classifiers helpful information, correlation and cointegration statistical tests were run. This compares daily user energy generation (usage transactions removed) directly to daily solar data. For each household, energy production transactions are taken and compared against all 100 solar data sets with correlation and cointegration. The time series are of the same length across the three year period of the original energy data. After a household's production is compared to each region's solar data, they are ranked using the correlation coefficient or cointegration t-statistic. The rank of the correct postcode of the household under analysis is taken as the score for each measure.

After all households are analysed the following values in table 7 were produced. The average ranks are out of 100 and show for correlation some predictive power, especially for the median and mode. Cointegration, however, was poor and no statistical link can be seen. The distribution of correlation ranks are shown in Figure 13 and show a heavy skew towards highly ranking the correct postcode of a household. However, the distribution has a large spread of results, making it far less useful in many cases. Overall this simple correlation of solar data to user energy production shows solar data should be expected to aid the classification as it has.

Table 7. **Correlation and cointegration**

Measure	Mean Rank	Stdev	Median	Mode
Correlation	27.86	29.21%	17	1
Cointegration	55.14	33.90%	64	76

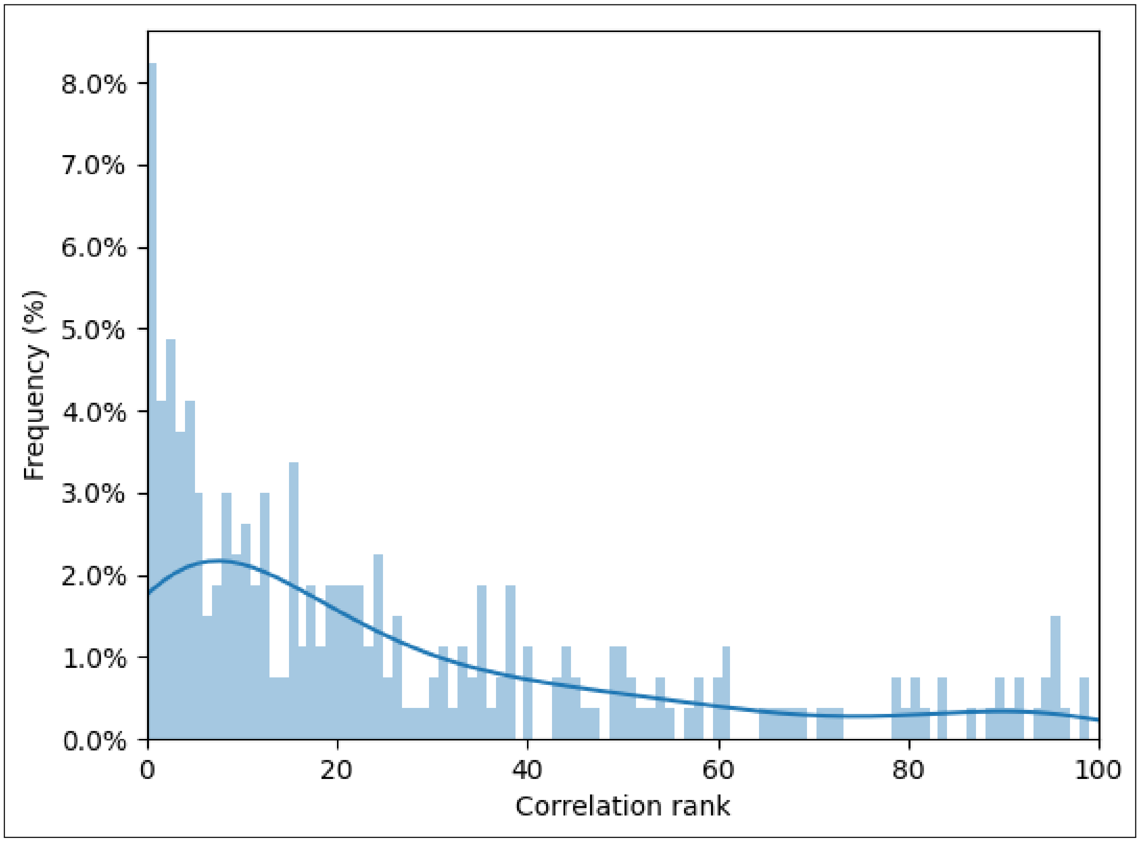


Fig. 13. Correlation distribution

Overall the machine learning models are consistently improved by including additional solar data which has some correlation to a household's energy production. This aids in both the cases where users take limited steps to protect their privacy and also the mixed ledger scenario. The next section will suggest and measure approaches for users to significantly reduce these high accuracies of prediction, without requiring the likely expensive process of using a new PK every transaction.

4.4. Objective Four - Obfuscation Techniques - Remaining

Progress and next steps

- Remaining: Ability to generate new datasets for the below tests.
- Remaining: Run new tests after varying PK numbers and mixing ledgers. Restricted to CNN and RFC.
- Remaining: Run new tests after applying timestamp obfuscation. Restricted to CNN and RFC.

5. CONCLUSION

To achieve the progress so far I have had to learn about blockchain implementation and privacy concepts. But more importantly delve into machine learning approaches beyond standard classification and clustering concepts learnt previously. This field of data manipulation and analysis is massive and exploring it is interesting and a good learning experience.

Overall machine learning models can quite accurately link user transactions from past blockchain data when users take limited steps to protect their privacy. More frequent transactions aid an attacker, as does using separated ledgers whether by users or postcodes. A classification accuracy of 85% has been achieved with a convolutional neural network model. Attacker success rates are consistently improved by including additional solar data which has some correlation to a household's energy production. This aids whether users protect their privacy or not. The final stages of the research to complete include adding half-hourly to the solar results and the final obfuscation techniques section. This will suggest and measure approaches for users to significantly reduce the high accuracies of prediction being seen.

6. REFERENCES

1. E. F. Jesus, V. R. L. Chicarino, C. V. N. de Albuquerque, and A. A. de A. Rocha, “A survey of how to use blockchain to secure internet of things and the stalker attack,” Secur. Commun. Networks pp. 1–27 (2018).
2. A. Dorri, C. Roulin, R. Jurdak, and S. S. Kanhere, “On the activity privacy of blockchain for iot,” (2019), pp. 258–261.
3. D. Ermilov, M. Panov, and Y. Yanovich, “Automatic bitcoin address clustering,” (2017), p. 461–466.
4. Ausgrid, “Solar home electricity data,” https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data (2014).
5. Bureau of Meteorology, “Climate data online,” http://www.bom.gov.au/climate/data/ (2020).
6. S. Nakamoto, “Bitcoin: A peer-to-peer electronic cash system,” (2008).
7. G. Wood, “Ethereum: A secure decentralised generalised transaction ledger,” https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Past-outage-data (2014).
8. M. Ferrag, M. Derdour, M. Mukherjee, A. Derhab, L. Maglaras, and H. Janicke, “Blockchain technologies for the internet of things: Research issues and challenges,” IEEE Internet Things J. **6**, 2188–2204 (2019).
9. D. Puthal, N. Malik, S. Mohanty, E. Kougianos, and G. Das, “Everything you wanted to know about the blockchain: Its promise, components, processes, and problems,” IEEE Consumer Electron. Mag. **7**, 6–14 (2018).
10. M. Swan, *Blockchain: Blueprint for a New Economy* (O’Reilly, Sebastopol, CA, USA, 2015), 1st ed.
11. T. Alladi, V. Chamola, J. Rodrigues, , and S. Kozlov, “Blockchain in smart grids: A review on different use cases,” Sensors (Switzerland) **19** (2019).
12. R. Bayindir, I. Colak, G. Fulli, and K. Demirtas, “Smart grid technologies and applications,” Renew. Sustain. Energy Rev. **66**, 499–516 (2016).
13. U. Ahsan and A. Bais, “Distributed big data management in smart grid,” 26th Wirel. Opt. Commun. Conf. pp. 1–6 (2017).
14. L. Cheng, N. Qi, F. Zhang, H. Kong, and X. Huang, “Energy internet: Concept and practice exploration,” (2017), p. 1–5.
15. M. Andoni, V. Robu, D. Flynn, S. Abram, D. Geach, D. Jenkins, P. McCallum, and A. Peacock, “Blockchain technology in the energy sector: A systematic review of challenges and opportunities,” Renew. Sustain. Energy Rev. **100**, 143–174 (2019).
16. V. Hassija, G. Bansal, V. Chamola, V. Saxena, and B. Sikdar, “Blockcom: A blockchain based commerce model for smart communities using auction mechanism,” (2019), p. 1–6.
17. G. Bansal, V. Hassija, V. Chamola, N. Kumar, and M. Guizani, “Smart stock exchange market: A secure predictive decentralised model,” (2019), p. 1–6.
18. J. Li, Z. Zhou, J. Wu, J. Li, S. Mumtaz, X. Lin, H. Gacanin, and S. Alotaibi, “Decentralized on-demand energy supply for blockchain in internet of things: A microgrids approach,” IEEE Transactions on Comput. Soc. Syst. **6**, 1395–1406 (2019).
19. B. Muhammad, J. Zhao, D. Niyato, L. Kwok-Yan, and X. Zhang, “Blockchain for future smart grid: A comprehensive survey,” IEEE Transactions on Comput. Soc. Syst. (2019).
20. P. Kumar, Y. Lin, G. Bai, A. Paverd, J. S. Dong, and A. Martin, “Smart grid metering networks: A survey on security, privacy and open research issues,” IEEE Commun. Surv. & Tutorials **21**, 2886–2927 (2019).
21. M. K. Khalilov and A. Levi, “A survey on anonymity and privacy in bitcoin-like digital cash systems,” IEEE Commun. Surv. & Tutorials **20**, 2543–2585 (2018).
22. M. Conti, S. Kumar, C. Lal, and S. Ruj, “A survey on security and privacy issues of bitcoin,” IEEE Commun. Surv. & Tutorials **20**, 2543–2585 (2018).
23. D. Ron and A. Shamir, “Quantitative analysis of the full bitcoin transaction graph,” (2013), pp. 6–24.
24. Z. Ghahramani, “Unsupervised learning,” Adv. lectures on machine learning **20**, 72–112 (2003).
25. H. H. S. Yin, K. Langenheldt, M. Harlev, R. R. Mukkamala, and R. Vatrapu, “Regulating cryptocurrencies: A supervised machine learning approach to de-anonymizing the bitcoin blockchain,” J. Manag. Inf. Syst. **36**, 37–73 (2019).
26. Y. Wang, G. Cao, S. Mao, and R. M. Nelms, “Analysis of solar generation and weather data in smart grid with simultaneous inference of nonlinear time series,” in *2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS),* (2015), pp. 600–605.
27. T. Khatib and W. Elmenreich, “A model for hourly solar radiation data generation from daily solar radiation data using a generalized regression artificial neural network,” Int. J. Photoenergy **2015**, 1–13 (2015).
28. Z. Guan, G. Si, X. Zhang, L. Wu, N. Guizani, X. Du, and Y. Ma, “Privacy-preserving and efficient aggregation based on blockchain for power grid communications in smart communities,” IEEE Commun. Mag. **56**, 82–88 (2018).
29. H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, “Deep learning for time series classification: a review,” Data Min Knowl Disc **33**, 917–963 (2019).
30. B. A, L. J, B. A, L. J, and K. E, “The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances,” Data Min. Knowl. Discov. **31**, 606–660 (2017).
31. J. Lines and A. Bagnall, “Time series classification with ensembles of elastic distance measures,” Data Min. Knowl. Discov. **29**, 565–592 (2015).
32. M. G. Baydogan, G. Runger, and E. Tuv, “A bag-of-features framework to classify time series,” IEEE Transactions on Pattern Analysis Mach. Intell. **35**, 2796–2802 (2013).
33. A. Bagnall, J. Lines, J. Hills, and A. Bostrom, “Time-series classification with cote: The collective of transformation-based ensembles,” IEEE Transactions on Knowl. Data Eng. **27**, 2522–2535 (2015).
34. Z. Wang, W. Yan, and T. Oates, “Time-series classification with cote: The collective of transformation-based ensembles,” Int. joint conference on neural networks p. 1578–1585 (2017).
35. A. Jović, K. Brkić, and N. Bogunović, “Decision tree ensembles in biomedical time-series classification,” (Springer, Berlin, Heidelberg, 2012), p. 917–963.
36. S. Bhandari, N. Bergmann, R. Jurdak, and B. Kusy, “Time series analysis for spatial node selection in environment monitoring sensor networks,” Sensors **18**, 11–27 (2017).
37. Open Power System Data, “Household data,” (2017).

7. APPENDIX A - ADDITIONAL TRANSACTION ANALYSIS GRAPHS

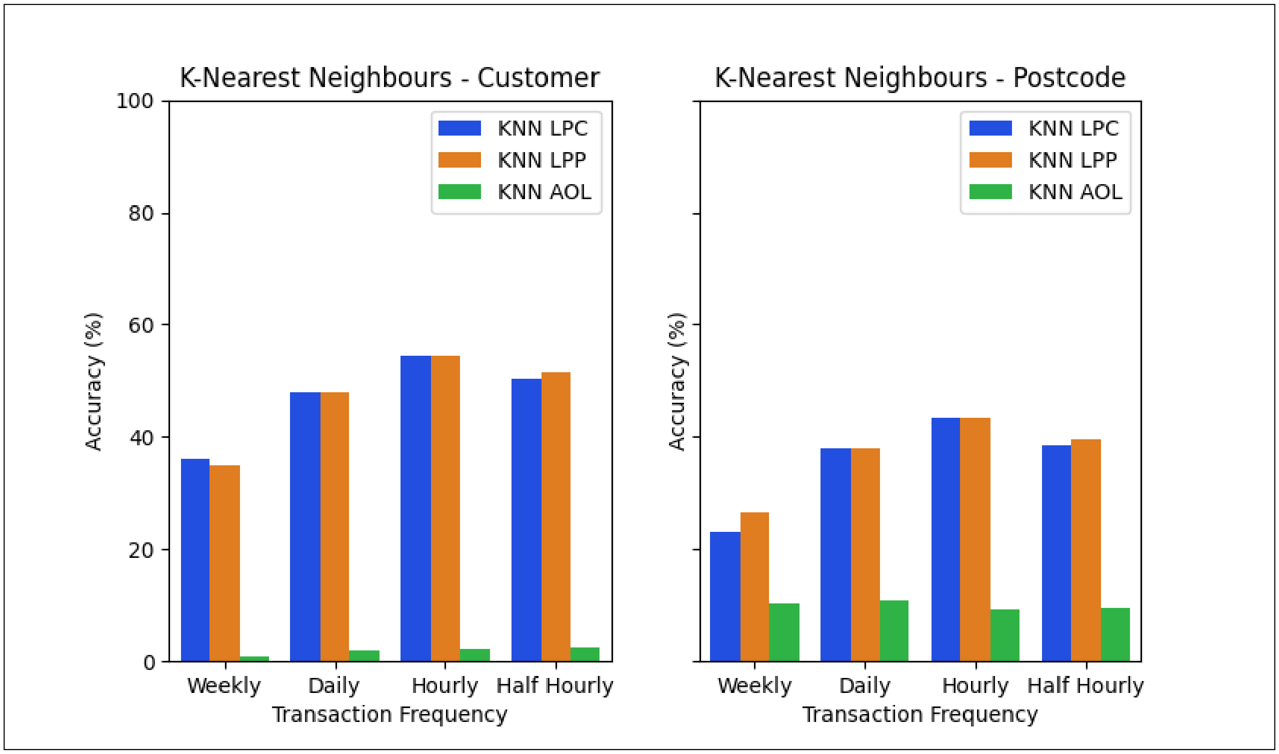


Fig. 14. Overall accuracy of KNN classification by transaction resolution and scenarios

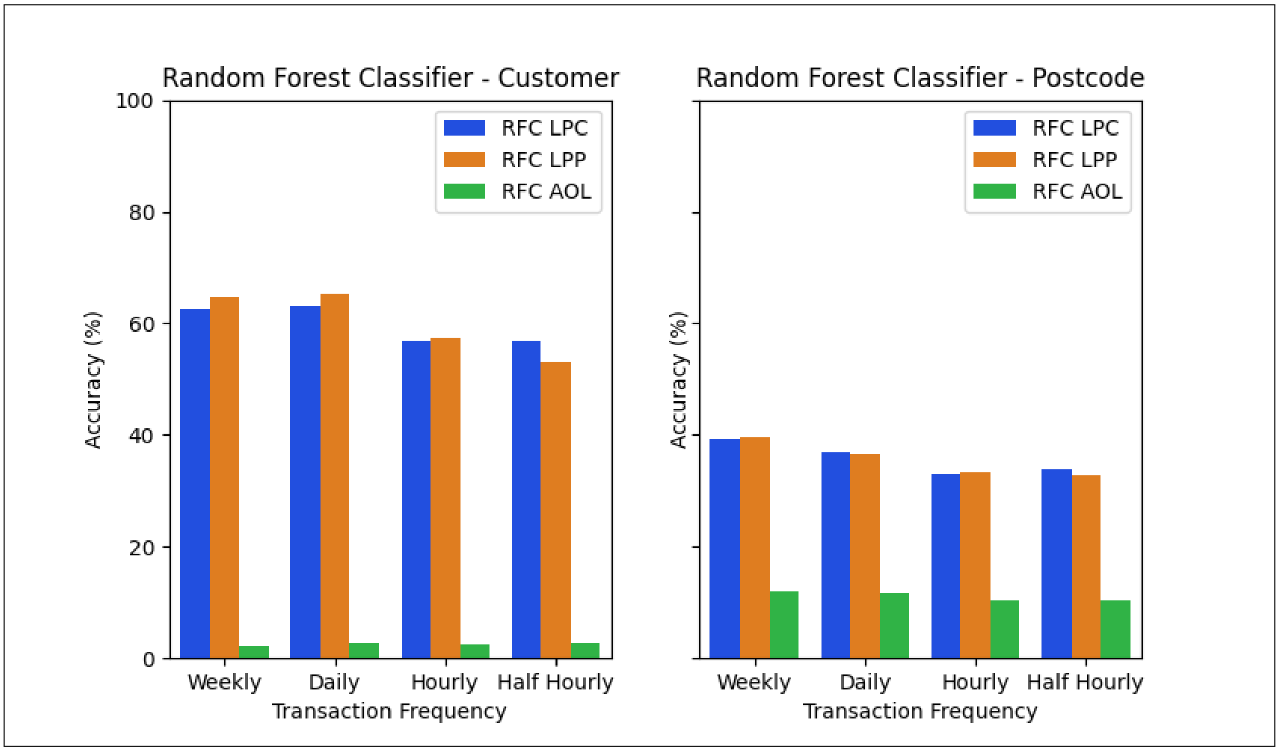


Fig. 15. Overall accuracy of RF classification by transaction resolution and scenarios

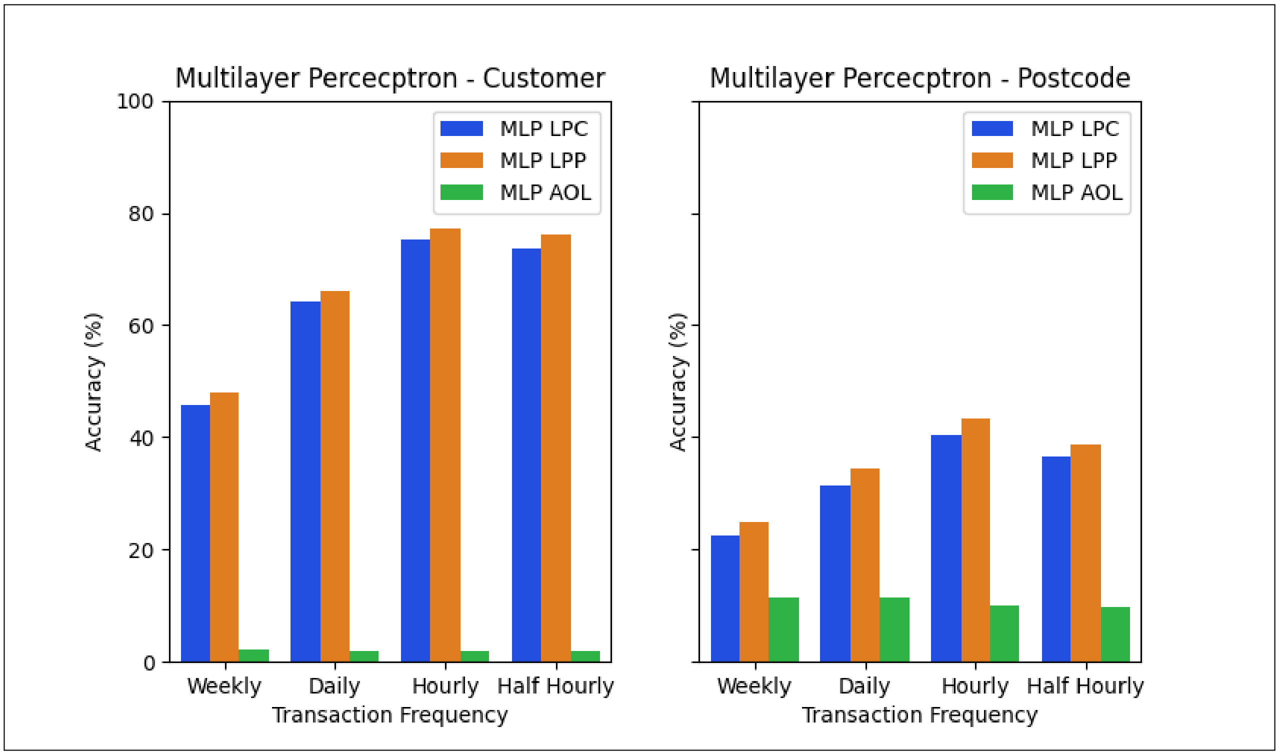


Fig. 16. Overall accuracy of MLP classification by transaction resolution and scenarios

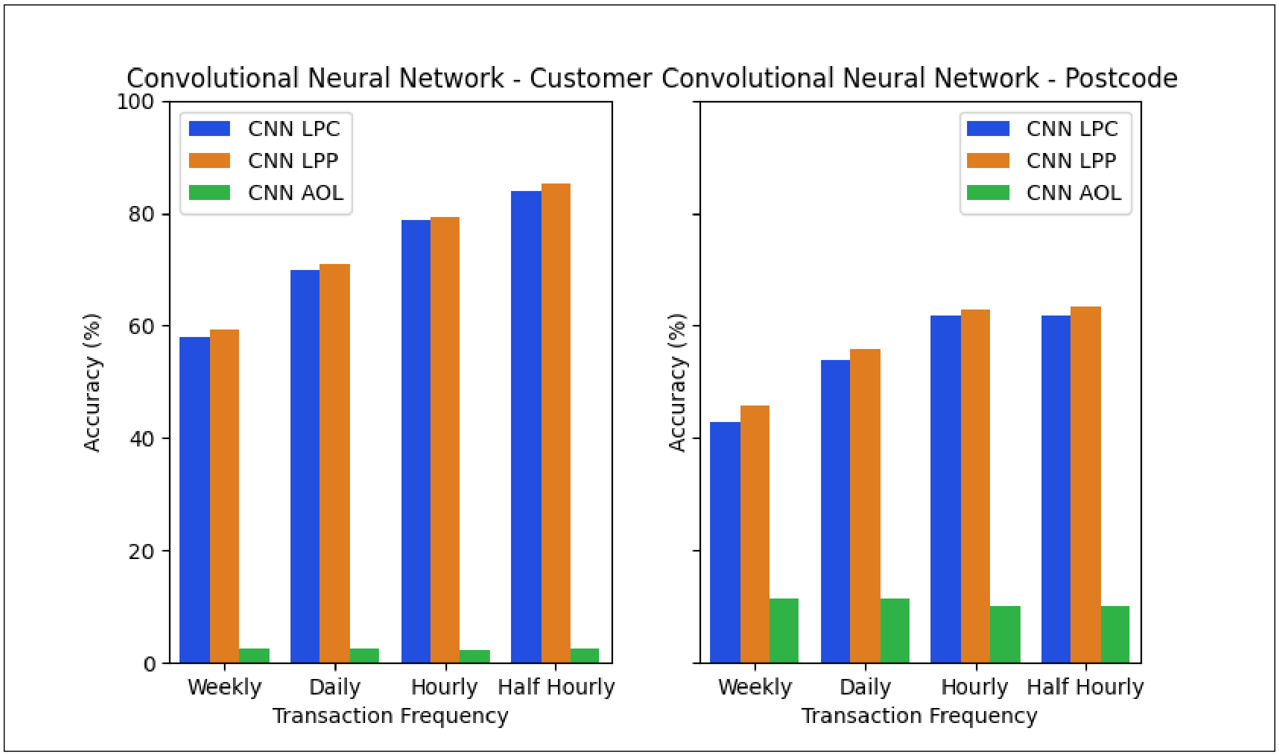
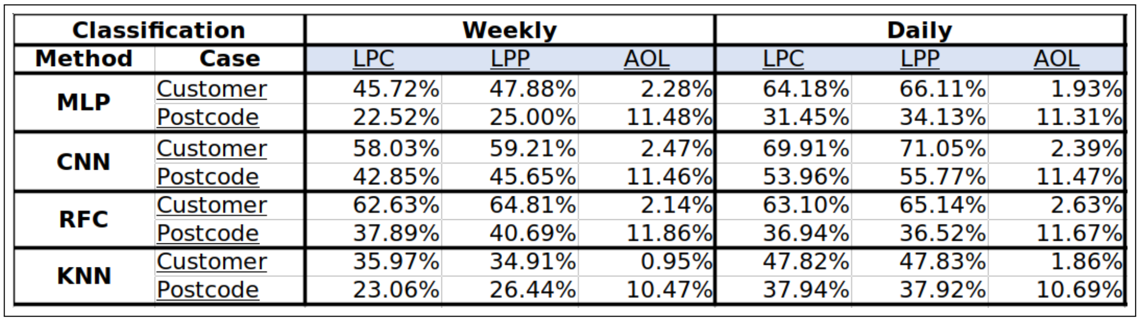


Fig. 17. Overall accuracy of CNN classification by transaction resolution and scenarios

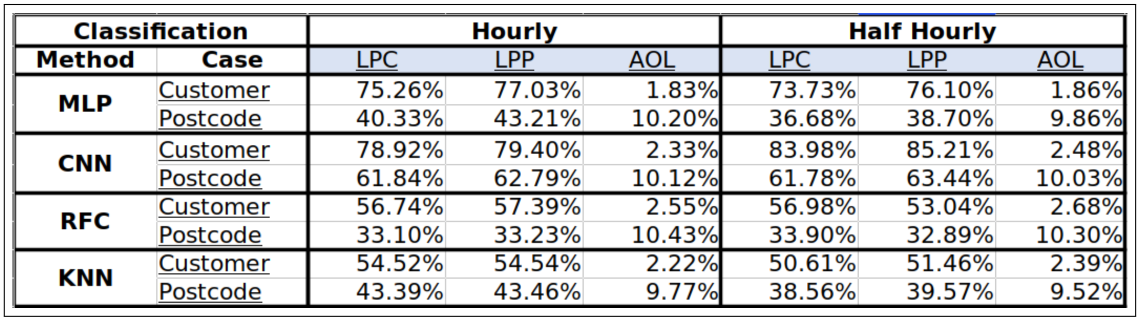
8. APPENDIX B - TABULATED RESULTS

Section 4.2 Objective 2 Tabulated Results



Classification		Weekly			Daily		
Method	Case	LPC	LPP	AOL	LPC	LPP	AOL
MLP	Customer	45.72%	47.88%	2.28%	64.18%	66.11%	1.93%
	Postcode	22.52%	25.00%	11.48%	31.45%	34.13%	11.31%
CNN	Customer	58.03%	59.21%	2.47%	69.91%	71.05%	2.39%
	Postcode	42.85%	45.65%	11.46%	53.96%	55.77%	11.47%
RFC	Customer	62.63%	64.81%	2.14%	63.10%	65.14%	2.63%
	Postcode	37.89%	40.69%	11.86%	36.94%	36.52%	11.67%
KNN	Customer	35.97%	34.91%	0.95%	47.82%	47.83%	1.86%
	Postcode	23.06%	26.44%	10.47%	37.94%	37.92%	10.69%

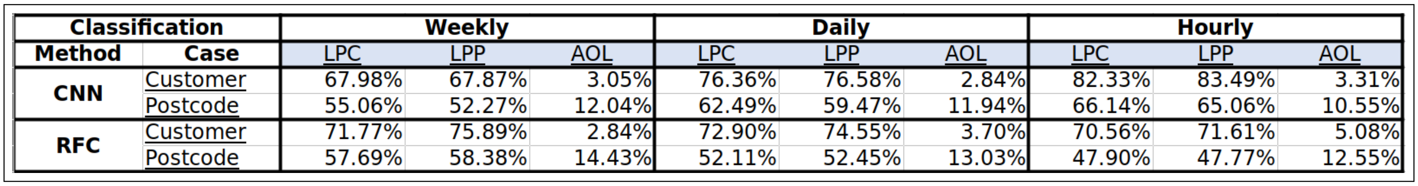
Fig. 18. Weekly and daily results for transaction classification



Classification		Hourly			Half Hourly		
Method	Case	LPC	LPP	AOL	LPC	LPP	AOL
MLP	Customer	75.26%	77.03%	1.83%	73.73%	76.10%	1.86%
	Postcode	40.33%	43.21%	10.20%	36.68%	38.70%	9.86%
CNN	Customer	78.92%	79.40%	2.33%	83.98%	85.21%	2.48%
	Postcode	61.84%	62.79%	10.12%	61.78%	63.44%	10.03%
RFC	Customer	56.74%	57.39%	2.55%	56.98%	53.04%	2.68%
	Postcode	33.10%	33.23%	10.43%	33.90%	32.89%	10.30%
KNN	Customer	54.52%	54.54%	2.22%	50.61%	51.46%	2.39%
	Postcode	43.39%	43.46%	9.77%	38.56%	39.57%	9.52%

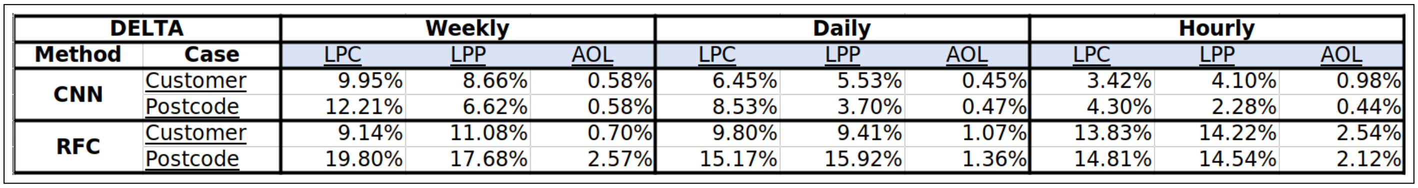
Fig. 19. Hourly and half-hourly results for transaction classification

Section 4.3 Objective 3 Tabulated Results



Classification		Weekly			Daily			Hourly		
Method	Case	LPC	LPP	AOL	LPC	LPP	AOL	LPC	LPP	AOL
CNN	Customer	67.98%	67.87%	3.05%	76.36%	76.58%	2.84%	82.33%	83.49%	3.31%
	Postcode	55.06%	52.27%	12.04%	62.49%	59.47%	11.94%	66.14%	65.06%	10.55%
RFC	Customer	71.77%	75.89%	2.84%	72.90%	74.55%	3.70%	70.56%	71.61%	5.08%
	Postcode	57.69%	58.38%	14.43%	52.11%	52.45%	13.03%	47.90%	47.77%	12.55%

Fig. 20. Results for classification with solar data



DELTA		Weekly			Daily			Hourly		
Method	Case	LPC	LPP	AOL	LPC	LPP	AOL	LPC	LPP	AOL
CNN	Customer	9.95%	8.66%	0.58%	6.45%	5.53%	0.45%	3.42%	4.10%	0.98%
	Postcode	12.21%	6.62%	0.58%	8.53%	3.70%	0.47%	4.30%	2.28%	0.44%
RFC	Customer	9.14%	11.08%	0.70%	9.80%	9.41%	1.07%	13.83%	14.22%	2.54%
	Postcode	19.80%	17.68%	2.57%	15.17%	15.92%	1.36%	14.81%	14.54%	2.12%

Fig. 21. Delta of initial results to results with solar data