Dissertation

* **Title**
  + Anomaly Detection on a Small Scale Network utilising machine learning techniques through a Multi-Layer Perception
* **Abstract**
  + Context
    - Machine learning is a newly evolving field of computer science, a branch of AI field of study that can be utilised for big data, giving it huge scope for future work and implementations into many complex computer systems as it has the facilities to analyse and monitor data sets that far exceed the human scope of analysis. IDS have been used for almost 30 years and as these new fields of study evolve in computer security the interconnectivity of these disciplines is moving from fiction to fact.
  + Aim
    - To replicate previously conducted studies in the design and production of a “Smart IDS” solution utilising machine learning as the basis for threat analysis and identification on a small scale network.
  + Method
    - The project will operate by using 4 Raspberry Pi’s as the target network each connected to a central router that will act as the gateway for the network with all data passing through it, this will be the home of the smart IDS. The Smart IDS itself will be constructed in Python utilising Theano and will be trained using the NSL KDD dataset so that the presented solution is able to correctly identify and classify the incoming attacks, present within the dataset. Upon reaching a satisfactory level of competence with the training module the MLP will then be deployed on the target network and tested in real time as similar attacks to those of the dataset will be launched at the network to replicate a real time attack, testing the MLP’s identification ability.
  + Results
    - The presented solution is capable of persistently, correctly identifying incoming attacks on the network with a positive classification percentage of (&&%) clearly demonstrating that the MLP has ‘taught’ itself the correct pattern recognition to identify the variants of incoming attacks on the network.

**Chapter 1 Introduction**

* Internet of Things
  1. Growth
     1. Internet of Things is the term used to categorize the newly emerging market of everyday devices that are becoming increasingly interconnected as a result of an internet connection. The purpose of this newly emerging technology is supposedly ease of life and continuing the trend of the last century of integrating technology into our day to day lives to ever increasing levels resulting in a more automated existence than mankind has ever experienced before.
  2. Untested/saturated marketplace
     1. Due to the rapid rise of this new technology the marketplace for these types of devices has become saturated with increasingly lower quality products and with companies wanting to be the first to produce these new pieces of kit there has been no regulation or standard set for the integrity of these devices. This has resulted in an overwhelmed market full of unsecure and fairly untested products that can leave the owners vulnerable to external threats.
  3. Security concerns
     1. As a result of the timescale of this interconnected plethora’s rise in popularity there have been serious shortcuts made on the security side of the products, this can be, in part, attributed to the surge of these types of devices being produced in China which are non-regulated and contain a far higher percentage of vulnerabilities than their Western counterparts. (ref on Chinas shit IoT)
     2. These devices can be at risk from a number of different attacks varying from Denial of Service attacks to man in the middle if the device contains potentially lucrative personal information about the owner which could be used to the attackers advantage.
* Intrusion Detection Systems
  1. An Intrusion Detection System is a security management system used by large scale corporations to monitor employee activities along with the company’s network to identify when attacks are occurring either internally or externally. The systems operate by having a profile for each user with their usual operations and locations for these and comparing this ‘normal behaviour’ against what the user is doing in real time, if there are anomalies or the actions seem out of the normal practice then the IDS has the power to halt the users actions along with reporting these actions to the Security Supervisor.
* Machine Learning
  1. MLP
     1. The field of Artificial Intelligence has in the last half century has moved from science fiction to reality, with the exponential growth of computing power and our increasing understanding of how our brain works this field of computing is rapidly approaching a generation defining breakthrough. While a stable sentient Artificial Intelligence is still on the horizon, a branch of this field of computing is gaining large scale academic following, Machine Learning. Instead of working towards sentiency this discipline focusses on pattern recognition in its most basic form, giving it great potential for big data or analysis of large datasets that humans would struggle greatly with.
     2. One of the central concepts of Machine Learning is an Artificial Neural Network (ANN) designed to be a virtual simulation of a neural network found in a human being, these generally operate by being fed a data stream that is processed through a hidden layer where complex functions are used against the data before being fed to the output layer. Machine learning operates by the program teaching itself the patterns from within the data that a human being couldn’t, as a result of this central pillar the systems have to be ‘taught’ what they are identifying.
  2. Types of Learning
     1. There are different ways that an ANN can be taught, supervised and unsupervised learning. Supervised learning works by the ANN being fed a labelled training dataset which will be used by the ANN to determine the pattern between the input and output of this data. Unsupervised learning uses an unlabelled data set and operate on a Self-Organising Map where the program learns to organize the data internally.
     2. Supervised Learning can be implemented in many different ways but one of the most popular and effective is backpropagation, data is paired with an input vector and a target vector once the input has been passed through the ANN it is compared against the target vector if these are different the Neural Network alters the weights it has used internally in an effort to reduce error rate of the output. This cyclic process is repeated until the error rate has reached a satisfactory level and the system can be thought of as having ‘taught itself the data’, hence the term Machine Learning.
     3. A type of ANN that utilises supervised learning techniques is a Multi-Layer Perception as these operate in the same way that a biological neural network would, with data being fed into it, processed in the ‘hidden layer’ where the complex functions and operations are performed before then passed to an output node where the error rate will be evaluated, if it is still not satisfactory the output is then passed back to the input where the entire process is repeated.

The specific question of this project has been to develop an IDS system that uses an MLP incorporating Supervised Learning techniques that can correctly identify when and what type of attack is being launched on a small scale IoT network composed on Raspberry Pi’s. This project will attempt to replicate the past work of (blah blah) in demonstrating that this form of IDS can indeed be developed and operate to a high level of accuracy.

(Potentially add more about what I am trying to achieve in the course of this project)

Set the scene  
Background to and purpose of the investigation  
Scope  
Include project aims/research question  
likely to be more focussed than in your proposal  
End with Overview of Remaining Chapters  
Don’t write this first. Wait to you know the whole story of your project   
Guide - (750-1000 words)

**Chapter 2 Literature review**

* Internet of Things
  + Rise (figures)
    - The exponential rise in popularity of IoT and the worldwide distribution of devices featuring exploits has created a low priority, global vulnerability epidemic with Cyber security experts predicting that the number of these devices will only continue to grow. (source)
    - On (date) there were (number) IoT devices connected worldwide whereas by (date) (source) stated that there were (number) devices connected, this clearly demonstrates the growth of this form of device and with the lacking regulation of this new form of technology a number of users are leaving themselves vulnerable. As with any new emerging technology there will always be illegitimate vendors pushing products that do contain vulnerabilities however when examining the rise of IoT devices in relation to other new technologies and their initial security issues there has never been a more widespread acceptance of faulty goods, (source) a huge portion of these interconnected devices have security risks.
  + Dangers
    - The introduction and rapid integration of these devices has not allowed the appropriate testing to be conducted on these products to ensure the safety of both the user and any user information that could be gained through exploiting one of these devices. The primary selling point of these devices is also their most dangerous asset, let us look an IoT thermostat, as this is connected to the users home Wi-Fi network they have the ability to put the heating on when they leave work so the house is warm for arrival but the heating is not wasted by being on all day. Environmentally beneficial and convenient, there is no denying that, however if this device has a hidden vulnerability residing within this could give an attacker initially access to the thermostat. What if there are other vulnerabilities within the users home network, an unpatched browser on a running PC for example, this could provide access into the home network and puts all the users personal data at risk, this is where the real danger of IoT devices lie.
  + Bot nets (ref)
    - Due to the high number of these devices that are connected to the web with vulnerabilities hidden within them, another serious threat that can be caused by these devices is that they become infected with a piece of malware called a botnet. A botnet works by infecting itself in as many hosts as possible, just like normal malware, once enough hosts have been infected the virus then secretly steals a portion of the infected host’s resources and processing power to use for malicious purposes. It is important to note that although these devices are not super computers and largely have very little processing power available to them, when a large number are combined they can produce an impressive amount of processing power which can be used for bitcoin mining or even password cracking.
* IDS
  + Intrusion Detection Systems have been a central pillar in large scale conglomerate security since their conception in the late 70s however the versions implemented at these early stages in IDS lifespan were far less sophisticated.
  + Due to the cost of electronic storage (R.Kremmerer) these earlier systems printed out their log files which would then require line by line analysis by one of the security personnel, making these early incarnations of IDS labour intensive and they provided no real time Intrusion Detection. As the technological revolution continued storage became cheaper and very basic programs were designed to analyse the data, real time protection was becoming more of a reality however the early versions of this software were very resource intensive and required a lot of computing power to move at any useful speed.
  + Intrusions into a system can usually be classed into one of two types, Anomaly or Misuses Intrusion, the latter is the term used for a positive result by the IDS meaning that the system has identified a pre-known attack type occurring on the system whereas Anomaly Intrusions represent unusual user activity that could be an indicator of a potentially unknown attack. Anomaly Intrusions are usually identified by comparing a standard user behaviour profile, what do they usually do and where, against the actions they are currently performing, if an anomaly between the two is detected this gets flagged.
  + Anomaly detections are by far the most resource intensive out of the two forms of intrusion as they require a vast quantity of data to be analysed and compared against a standard behaviour database for the users of the system, due to the large quantity of data and one of the key fundamentals of IDS being pattern recognition the implementation of machine learning into these security systems began moving from theory to reality as the study of machine learning blossomed.
* Machine Learning
  + Machine learning is one of the youngest disciplines in the field of computing boasting nearly endless possibilities for its uses in our increasingly technologically based world. As (source) states one of the oldest fundamental problems in computing has been transferring the rationale based problem solving and critical thinking that human beings have developed over a lifetime, into a completely static electronic learning environment. For example if humans see a dog they will associate the animal with a lead or collar as we have spent our entire lives connecting these items together whereas a computer has no pre-existing knowledge base to work from, presenting the problem described. Machine learning and the study of artificial intelligence are the first steps towards computers developing the critical thinking and association we take for granted. Two of the first building blocks for the development of this are the study of image and pattern recognition as these can be thought of as the first step towards automated association.
  + Different kinds of systems
  + Different architectures
  + Utilisations as IDS examples using ANN
* Previous research done in both fields

A record of important contributions to your field of work by others and a critical evaluation of the work to date.

A more in-depth and focussed discussion than in your proposal

Focus on actual technologies now being used

Current state of play

Start with short introduction of what chapter is about

References to support arguments

Lead reader to same position as you

End with Summary of significant points leading in to your methodology

Guide (1500-2000 Words)

**Chapter 3 Methodology - Practical Aspects**

* How the victim network will be configured
  + How the network data will be captured – Wireshark
* How the program itself is constructed, what utility is used to build/operate it
  + Anaconda – coded in python
  + Due to the MLP using Theano, a python directory for the machine learning section
    - Theano has access to the algorithms required for this part of the artefact
* How the MLP IDS is designed, how does it work – brief explanation of the workings of MLP
  + Into more detail and actually explain
* Training of the system
  + NSL KDD – different variations which will be used during training to get the model to a state of readiness
  + What the program will identify
  + How will false positives be dealt with
* Testing Phase
  + Program will be implemented into the victim network and initialised
    - Traffic will be supplied using python scripts to provide diversity to the dataset the model was previously trained on
    - Data from the network will be recorded with Wireshark and fed to the program and used as the dataset for the current run of the machine
  + DDOS and Man in the middle attacks will then be carried out on the network to see if the program can identify threats in real time
  + Recursive cycle of development so the testing phase will be carried out multiple times

It should be possible for a worker in your field to repeat your method and results

A description of how you carried out your project

Justify why you have done what you have done

Guide – 2500 words

**Chapter 4 Results**

A factual presentation of your results which relate to the project aim

A description of the completed software/hardware and analysis along with test/evaluations/analysis results.

Suitably presented:

Tables, Charts

Statistics

Illustrations

If too many use appendices

Put raw data in the appendices

Don’t dwell on discussion of issues. Save to Discussion chapter

Guide – 500 -1000 words

**Chapter 5 Discussion**

Evaluate your findings

Comment on their significance in relation to previous work on the same topic

Refer back to your literature review where appropriate

Use the key performance indicators outlined in your proposal if appropriate to aid your evaluation, referring back to initial project requirements

Guide – 2000 - 2500 words

**Chapter 6 Conclusion and recommendations/future work**

What conclusions can you draw from your investigation?

What are the implications of what you have discovered?

How might further work in this area be continued?

Guide -750 - 1000 words