Outlier Detection in Virtual Machine Resource Data

A dissertation submitted in partial fulfilment of

the requirements for the degree of

BACHELOR OF ENGINEERING in Computer Science

in

The Queen's University of Belfast

by

Liam Reid

3rd May 2022

SCHOOL OF ELECTRONICS, ELECTRICAL ENGINEERING and COMPUTER SCIENCE

CSC3002 – COMPUTER SCIENCE PROJECT

Dissertation Cover Sheet

A signed and completed cover sheet must accompany the submission of the Software Engineering dissertation submitted for assessment.

Work submitted without a cover sheet will **NOT** be marked.

|  |  |  |  |
| --- | --- | --- | --- |
| Student Name: | Liam Reid | Student Number: | 40231992 |
| Project Title: | Outlier Detection in Virtual Machine Resource Data. | | |
| Supervisor: | Dr Son. T. Mai | | |

Declaration of Academic Integrity

Before submitting your dissertation please check that the submission:

1. Has a full bibliography attached laid out according to the guidelines specified in the Student Project Handbook
2. Contains full acknowledgement of all secondary sources used (paper-based and electronic)
3. Does not exceed the specified page limit
4. Is clearly presented and proof-read
5. Is submitted on, or before, the specified or agreed due date. Late submissions will only be accepted in exceptional circumstances or where a deferment has been granted in advance.

**By submitting your dissertation you declare that you have completed the tutorial on plagiarism at** [**http://www.qub.ac.uk/cite2write/introduction5.html**](http://www.qub.ac.uk/cite2write/introduction5.html) **and are aware that it is an academic offence to plagiarise. You declare that the submission is your own original work. No part of it has been submitted for any other assignment and you have acknowledged all written and electronic sources used.**

1. If selected as an exemplar, I agree to allow my dissertation to be used as a sample for future students. (Please delete this if you do not agree.)

|  |  |  |  |
| --- | --- | --- | --- |
| *Student’s signature* | Liam Reid | *Date of submission* | 03/05/22 |

Acknowledgements

To those who have helped the author during the project and the preparation of the dissertation and to anybody who has given financial support.

Abstract

A summary (100 words) which provides an outline of the subject matter and the results, findings and/or conclusions of the dissertation.

Contents

A complete list of chapters, sections, appendices etc. with page numbers.

Main Text (see below)

The main body of the dissertation as described below organised as a sequence of chapters each normally containing several sections. The main text should not normally exceed 45 pages (it may be less).

References

A list of references to documents (books, papers, web pages etc.) which are referred to in the main body of the text. Use the IEEE citation style as detailed here <https://www.ieee.org/documents/ieeecitationref.pdf>. There is some guidance on referencing at <http://www.qub.ac.uk/cite2write/home.html>.

**The first citation should be the URL to the software code repository which should contain the code and any other resource required to run the software.**

Appendices

These should include as appropriate:

(a) A User manual giving details on how to use the software, including details of input data, output formats and error messages.

(b) Test results, if appropriate.

(c) Other information which is not convenient or appropriate to include in the main body of the dissertation.

The Main Text

1. Introduction and Problem Area
   1. Introduction

Cloud computing, although not fundamentally new, has become one of the fastest growing technologies in the 21st century [1]. Applications and services deployed on cloud-based platforms are becoming increasingly complex with the rise in popularity. Consequently, cloud engineers are constantly working to make their platforms robust and failsafe, however problems with cloud services are still prevalent as we move into 2022 [2].

On October 4th, 2021, Facebook, WhatsApp, and Instagram suffered an outage due to a problem with their shared cloud infrastructure. The effects were felt by its billions of users globally, claiming they felt “discomfort and displeasure” during the 6-hour period [2][3]. Facebook shares fell 5.4% and loses in revenue were estimated to be around $99.75 million [4]. Disasters like this one can have huge legal and financial implications for an organisation, therefore it is essential that cloud systems are constantly maintained and monitored.

Maintaining cloud systems within an organisation is the responsibility of the Cloud Systems Administrator. A Cloud Systems Administrator deploys services, monitors and analyses cloud resource performance and resolves any issues reported. This role has become so important in recent years that organisations dedicate entire teams of Cloud System Administrators to manage their cloud infrastructure [5].

With the increased complexity of this role, a software solution has been developed as part of this project in order to assist Cloud Systems Administrators in the monitoring of cloud services. This solution provides a platform to detect discrepancies in cloud resource data using newly implemented outlier detection techniques. The detection can run in real time to monitor CPU usage and identify outliers. The solution is also a platform for performing experiments on new outlier detection techniques with a variety of datasets. The proposed solution will aid Cloud Systems Administrators in performing their role and automate the process of detecting discrepancies in cloud resource usage.

* 1. Cloud Computing in 2022

Over 90% of organisations use some type of cloud service for a variety of different applications [6]. “Data backup, disaster recovery, email, virtual desktops, software development and testing, big data analytics, and customer-facing web applications [7]” are among those listed by Amazon Web Services (AWS).

In recent years, millions of people around the world have been forced to work from home because of the coronavirus crises\*\* [8]. Remote working was forced upon many unprepared organisations and their workforces. Consequently, the use of cloud-based virtualisation technology was utilised to provide staff with the resources necessary to carry out their roles remotely [9]. BYOD (Bring Your Own Device), paired with a VDI (Virtual Desktop Interface), provided employees with an interface to their organisation’s network, applications and resources [10], [11]. Therefore, with the aid of cloud technology, organisations were able to operate as normal through the height of the pandemic.

* 1. Risks / Mitigation of Hosting Cloud Resources

Resource exhaustion is a major risk to a VM host server, where software using a server will exhaust its resources and effect the availability of other Virtual Machine’s sharing the same resources [12]. Resource exhaustion may occur for several reasons, for example, an infinite loop that prevents a process from terminating or a cyber-attack that results in a server’s resources being misused. If availability is impacted, a systems software may fail and introduce vulnerabilities to the system [13]. An effective Outlier Detection algorithm paired with a monitoring dashboard can reduce the risk to availability by notifying the user of a Virtual Machine on a server exhausting more resources than it normally needs.

Physical resources for Virtual Machines are managed by software called the hypervisor. The hypervisor is responsible for provisioning new VMs and distributing the resources where necessary. Hypervisor security can be breached by ‘various malicious attacks.’ “Session hijacking, man-in-the middle attack, flooding attack (and) Malware-Injection attack(s)” can allow malicious intruders unauthorized access to a server’s resources. In session hijacking, a hacker will steal a legitimate user’s session ID when it is generated after login. Once the session is hijacked, the hacker can copy the VM and gain access to all the data. A hacker can use a man-in-the-middle attack to intercept data and gain access to a VM. A hacker can perform a Malware Injection attack by inserting malicious code into an application. This makes the resources, applications, and data on the VM vulnerable

[14]. These events are liable to happen often and when they do, they can go undetected. An Outlier Detector would mitigate the damage done to a system by notifying the user of any suspicious/anomalous behaviours.

Misuse or unauthorized access of a VM management console can be a serious risk. A hacker could gain access to a VM management console using techniques like phishing, malware attacks, brute force attacks or by learning login credentials from a system or website breach. Not just hackers, but an insider from an organisation with authorised access to VM management tools can also be a risk. With access to a VM management console, a person could cause a lot of damage by deprovisioning VMs running important processes. They could also provision additional VMs that are not needed, this would result in a huge cost for the organisation as resources are scaled to meet demand and if resources restrictions are set, availability of resources is decreased for other VMs running important processes [15]. These changes could easily be detected by a VM Outlier Detector dashboard and the risk and cost to the organisation can be reduced.

VM sprawl is the process where a VM is duplicated then forgotten about. These unmanaged VMs can operate on a network for weeks or even months without being detected. Without proper monitoring, these VMs can miss important security patches. As a result, there is an increased risk to security. There are methods to prevent VM sprawl, processes to govern VM lifecycle management using automated scripts can reduce the risk of VM sprawl, but this comes at a cost [12]. A dashboard monitoring Virtual Machine data in real time could quickly identify these forgotten VMs and notify the user.

[1] P.K. Paul, A. Kumar. and M. Ghosh. "Cloud Computing: the 21st Century Friend for Virtualization." *About Mewar University* (2012). Available: [Paper Title (use style: paper title) (researchgate.net)](https://www.researchgate.net/profile/Ajay-Kumar-80/publication/325083266_Cloud_Computing_the_21st_century_friend_for_virtualization/links/60eeb60f16f9f3130080302c/Cloud-Computing-the-21st-century-friend-for-virtualization.pdf)

[2] Cast AI. (2022, Jan. 11) “The cloud in 2021: 21 game-changing outages, security issues, and highlights” [Online]. Available: <https://cast.ai/blog/the-cloud-in-2021-21-game-changing-outages-security-issues-and-highlights/>

[3] N.M. Shousha. and L.R. Abdelgawad. (2021) "DOWN IN MINUTES, OUT FOR SIX HOURS: A BRIEF REPORT ON FEELINGS DURING THE OUTAGE OF WHATSAPP, INSTAGRAM, AND FACEBOOK." *British Journal of Psychology Research Vol.9, No.2, pp. 38-44.* Available: [DOWN IN MINUTES, OUT FOR SIX HOURS](https://www.researchgate.net/profile/Nayera-Shousha/publication/356193254_Down_in_Minutes_out_for_six_Hours_A_Brief_report_on_feelings_during_the_outage_of_WhatsApp_Instagram_and_Facebook/links/619203fe61f09877209a3736/Down-in-Minutes-out-for-six-Hours-A-Brief-report-on-feelings-during-the-outage-of-WhatsApp-Instagram-and-Facebook.pdf)

[4] The Guardian. (2021, Oct. 10) Impact of WhatsApp, Facebook and Instagram outage on businesses [Online]. Available: <https://guardian.ng/technology/social-media/impact-of-whatsapp-facebook-and-instagram-outage-on-businesses/> (Accessed 28th March 2022)

[5] J. Frank. (2019, Jul. 29) Your Next Move: Cloud Systems Administrator[Online]. Available: <https://www.comptia.org/blog/your-next-move-cloud-systems-administrator>

[6] 451 Research. (2017) Voice of the Enterprise: Cloud Transformation survey [Online] Available: <https://451research.com/images/Marketing/press_releases/Pre_Re-Invent_2018_press_release_final_11_22.pdf>

[7] AWS. (n.d.) Who is using cloud computing*?*[Online]. Available: <https://aws.amazon.com/what-is-cloud-computing/>

[8] S. Soares, F. Bonnet, J. Berg. (2021, Apr. 25) Working from home during the COVID-19 pandemic: Updating global estimates using household survey data[Online]. Available: <https://voxeu.org/article/working-home-during-covid-19-pandemic-updated-estimates>

[9] D. Anderson, C. Kelliher (2020), "Enforced remote working and the work-life interface during lockdown", *Gender in Management, Vol. 35 No. 7/8, pp. 677-683*. Available: <https://doi.org/10.1108/GM-07-2020-0224>

[10] P.J. Cho, et al. "Demographic Imbalances Resulting from Bring-Your-Own-Device Study Design." *JMIR mHealth and uHealth* (2021). Available: <https://s3.ca-central-1.amazonaws.com/assets.jmir.org/assets/preprints/preprint-29510-accepted.pdf>

[11] T. Mangan (2020, Mar. 30) The Tech Supporting Remote Workers[Online]. Available: <https://www.nutanix.com/theforecastbynutanix/technology/modern-remote-workforce-is-powered-by-virtual-desktop-technology>

[12] A. Chaudhuri. (2015). "Best Practices for Mitigating Risks in Virtualized Environments" - *Cloud Security Alliance, USA.* Available: <https://downloads.cloudsecurityalliance.org/whitepapers/Best_Practices_for%20_Mitigating_Risks_Virtual_Environments_April2015_4-1-15_GLM5.pdf>

[13] EBA/GL. (2017) Guidelines on ICT Risk Assessment under the Supervisory Review and Evaluation process *(SREP) (Page 11). Final Report on ICT Risk Assessment Guidelines.* [Online] Available: [EBA BS 2017 131 (Final Guidelines on ICT Risk Assessment under SREP) (europa.eu)](https://www.eba.europa.eu/sites/default/documents/files/documents/10180/1841624/ef88884a-2f04-48a1-8208-3b8c85b2f69a/Final%20Guidelines%20on%20ICT%20Risk%20Assessment%20under%20SREP%20%28EBA-GL-2017-05%29.pdf?retry=1) (Accessed: 4 January 2022).

[14] N. Arya., M Gidwani., S.K. Gupta. (2013) “Hypervisor Security - A Major Concern” *International Journal of Information and Computation Technology. Volume 3, Number 6, pp. 533-538.* Available: <https://www.ripublication.com/irph/ijict_spl/08_ijictv3n6spl.pdf>

[15] T. Morrow. (2018) 12 Risks, Threats, & Vulnerabilities in Moving to the Cloud. *Carnegie Mellon University's Software Engineering Institute Blog* [Online]. Available: [http://insights.sei.cmu.edu/blog/12-risks-threats-vulnerabilities-in-moving-to-the-cloud/](http://insights.sei.cmu.edu/blog/12-risks-threats-vulnerabilities-in-moving-to-the-cloud/%20) (Accessed March 28, 2022)

1. Solution Description System Requirements and Specification
   1. Development of an Effective Outlier Detector

The main technique developed to detect outliers is an ensemble of weak classifiers that work together and vote to determine a final classification [1]. The key elements considered when implementing this technique are described below.

[1] T.G. Dietterich. (2002)"Ensemble learning." *The handbook of brain theory and neural networks 2.1* *110-125.* Available: <https://courses.cs.washington.edu/courses/cse446/12wi/tgd-ensembles.pdf>

Selecting an Effective Technique for Detecting Outliers

An outlier is described as “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism [1]”

Outlier detection is used in many different fields for a variety of reasons. For example, banks may use outlier detection to detect fraud in abnormal spending patterns, a hospital to detect irregularities with a patient’s heartbeat, or in sports where a team evaluates player performance to determine outstanding attributes [2].

With the increased use of cloud technology, competition for resources on shared networks is growing, and such networks can become overloaded, resulting in increased downtime and a decrease in availability and reliability. Therefore, an effective solution has been proposed for detecting outliers in cloud resource data [3].

There are many different techniques for detecting outliers with statistical, nearest neighbour, clustering and classification based detection being among the most popular [4]. The technique proposed as part of this project uses a combination of statistical and classification based outlier detection which is especially effective against time series data. A time series can be defined as “a sequence of data points that occur in successive order over some period of time” [5]. The technique proposed is an ensemble of ‘weak’ classifiers that work quickly using statistical methods to examine the behaviour of the data over time, classifying data as either an outlier or an inlier. This technique accounts for the idea of concept drift, whereby the relationship between data points and time change throughout the series [6], and makes the classification quickly so that real-time outlier detection is possible. This is a key feature to this detector as it provides Cloud Systems Administrators the ability to detect discrepancies with performance and quickly diagnose issues with a cloud service.

Choosing a Learning Style

The problem with real world datasets that they are unlabelled (outliers are not known) [7]. The software developed will need to be able to detect outliers without labelled data. Therefore, it is important that different learning styles are used. There are three learning styles in which traditional classifiers are trained with.

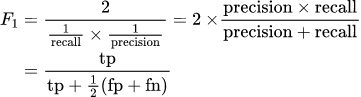
* Unsupervised
  + This learning style has access to data which contains outliers, but the data is not labelled. This means that the unsupervised model must work on its own to figure out what is an outlier and what is not.
* Supervised
  + Has access to labelled data containing both outliers and inliers. These models tend to be more accurate than unsupervised but depend on data being labelled correctly by a human.
* Semi-Supervised
  + Access to only a small amount of labelled data to train the model. This is useful when it’s only possible to label a limited amount because the dataset is so large and intricate.

[8]

Research shows that the supervised learning method produces the best results, but this cannot be used for real-world data since it is unlabelled. TALK ABOUT REAL TIME STREAM

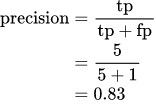
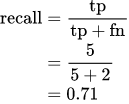
Outlier Detection Performance

To test the effectiveness of the technique implemented, and to compare it to traditional techniques, it must be tested on real world datasets. A comparison can be made by scoring the classifiers on their ability to detect outliers. Scoring the detection techniques can be achieved by generating an ‘f1’ score. An f1 score can be calculated using the following algorithm [9].



Equation: Calculating F1 Score [10].

Precision and recall are needed to calculate the f1 score, they can be calculated using the following algorithms.

Equations: Calculating Precision and Recall [10].

The software implementation can calculate the true positive, false positive and false negative values automatically. To calculate these values, the classifiers are used on labelled data sets (outliers are identified but not known to the classifier).

Ensemble Voting

For the detectors to work together they must vote on a classification for a specific data point. Two voting systems were proposed.

The first, ‘Majority Classification’, makes a prediction based on the individual classifications made by the detectors in the ensemble. The classification with the most occurrences therefore wins the vote. This method of voting runs quickly and experiments show it produces accurate results.

The second voting system, ‘Combined Confidence’, extracts a confidence value from each detection made by the detectors and combines them to generate a final confidence. Confidence is calculated from the distance a data point is from the thresholds of each detector. Results show this technique produces good f1 scores, however it is slower than the aforementioned ‘Majority Classification’ voting system.

NAB SCORING

[1] D.M. Hawkins. (1980). Introduction. In: Identification of Outliers. *Monographs on Applied Probability and Statistics. Springer, Dordrecht.* Available: <https://doi.org/10.1007/978-94-015-3994-4_1>

[2] H.P. Kriegel., P Kroger., A. Zimek. (2010) “Outlier Detection Techniques” *The 2010 SIAM International Conference on Data Mining Slide 6-7*. Available: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.183.1847&rep=rep1&type=pdf>

[3] H. Zhang., S. Chen., J. Liu., Z. Zhou., T. Wu (2017, Nov. 8) An incremental anomaly detection model for virtual machines. Available: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187488>

[4] V. Chandola. A. Banerjee., V Kumar. (2009, Jul) Anomaly Detection: A Survey. *ACM Computing Surveys, Vol. 41, No. 3, Article 15.5.* Available: <https://dl.acm.org/doi/epdf/10.1145/1541880.1541882>

[5] A. Hayes. (2021, Apr. 24) What Is a Time Series? [Online]. Available: <https://www.investopedia.com/terms/t/timeseries.asp>

[6] J. Gama., I. Žliobaitė., A. Bifet., M. Pechenizkiy., and A. Bouchachia. (2014) A survey on concept drift adaptation. *ACM Comput. Surv. 46, 4, Article 44 (April 2014), 37 pages*. Available: <https://doi.org/10.1145/2523813>

[7] P. Peng., T. Xiang., Y. Wang., M. Pontil., S. Gong., T. Huang., Y. Tian. (2016) Unsupervised Cross-Dataset Transfer Learning for Person Re-Identification *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1306-1315.* Available: <https://openaccess.thecvf.com/content_cvpr_2016/html/Peng_Unsupervised_Cross-Dataset_Transfer_CVPR_2016_paper.html>

[8] J. Delua. (2021, Mar. 12) Supervised vs. Unsupervised Learning: What’s the Difference? *IBM Analytics, Data Science/Machine Learning* [Online]. Available: <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning>

[9] Y. Sasaki. (2007). The truth of the F-measure. Teach Tutor Mater. Available: <https://www.researchgate.net/publication/268185911_The_truth_of_the_F-measure>

[10] T. Wood. (n.d.) Machine Learning Glossary and Terms – F-Score [Online]. Available: <https://deepai.org/machine-learning-glossary-and-terms/f-score>

* 1. Developing an Effective Software Solution

The software solution comes in the form of monitoring software. This software is used to detect outliers using a user specified detector on a user specified dataset. Users can monitor data in real time with outlier detection techniques applied or perform experiments on previously labelled datasets to determine which dataset works best for their data. This section specifies the key elements considered when developing this solution.

Programmatically Detecting Outliers

The software solution implemented contains programmatic implementations of various traditional outlier detection techniques. Techniques implemented have been programmed to take input, and produce a classification on data points in the same fashion. In doing so, clean code is achieved [1]. Less code is needed to display and score detection data since using this generic style makes code near the front end re-usable. Fig. 1 describes the outlier detection process. The ‘Data’ being passed to the outlier detector contains the time series data and any parameters needed by the detector to perform the detection. The output of this process is a data frame containing the outliers detected.

Programmatically Scoring Outliers

Some of the datasets available to run through this system are labelled meaning true positive, false positive, true negative and false negative values can be determined and the detector can be scored. As mentioned in section 2.1, precision, recall and f1 are used to generate a score for the detection. Accuracy is also calculated using this software solution [2]. Outlier detection data that has been processed by the detection is fed to the scoring system with the true outlier labels and the metrics are calculated.

Configuration of Detectors and Datasets

The software has been developed to achieve the following functionality.

* Monitor and detect outliers in cloud CPU usage in real time.
* Perform detection techniques on time series datasets and evaluate the techniques used.

To monitor and detect outliers in real time, the software reads CPU usage data from a location that is constantly updated. The CPU usage is plotted on a graph against time and outliers detected are marked on the graph and displayed to the user. The user can select a detector to perform detection in real time and switch to a different feed with CPU usage data for different VMs.

The software contains the functionality to perform experiments and generate scores for detectors. There are many options to assist the user in performing detection on a specified dataset. The user can select a detector from a dropdown box containing all detectors implemented in the system. New detectors can be copied to the projects code base, a few small tweaks in the detection script and configuration can allow the user to use a newly implemented detector with the monitoring software. The user can select a dataset from a dropdown containing all datasets that are available in the project resources. If a new dataset needs to be added, it is simply copied to the project resources and the configuration (containing the dataset name) is updated. The dataset is then available for use by the user.

This is made possible through the use of clean code practices throughout the code base. The monitoring software is highly configurable allowing for many combinations of detectors and datasets. This provides a thorough experimentation platform for examining detectors with multiple different time series datasets.

Representing Data

Data readability is greatly improved by visualising graphs, with this in mind, the main component of the real time detection is an animated graph that updates in real time [3]. Users can visualise the stream of data over time with outliers detected marked by red circles. Additionally, a pie chart depicting the availability of the CPU is displayed to the user. The pie chart is updated in real time and represents CPU usage in an intuitive format.

In the experimental space of this software, graphs are the key component in representing the detector classifications. Time series data is plotted, true positive classifications are represented by green dots, false positive by red dots and false negative by black. This graph, along with text containing accuracy, recall, precision and an f1 score, provide a user with all the data they need to evaluate the detection.

Storing Data

Data is stored so that generated detection can be recalled later. This saves time when performing experiments since some detection techniques take several minutes to perform. A database has been implemented to store the detection data. It has been designed using common practices such as normalisation. Tables within the database do not contain data that can be derived [8]. For example, accuracy, precision, recall and f1 are not stored since they can be derived from the true positive, false negative, false positive and true negative values. There are many combinations of detectors and datasets, so lots of data may need to be stored, the database provides a scalable solution for this applications needs.

Usability

Usability is described in part 11 of the ISO 9241 as “The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.”[4]. Good UI design is crucial to the success of any client facing software system so it is important to develop a UI that is customer focused [5]. The UI implemented with this software system is minimalistic. It contains a limited amount of components consisting of dropdown boxes and graphs. For monitoring software, a bombardment of information would take away from the focus of the application which is detecting outliers. There is no need for the software to display lengthy text to the user, the primary focus is to detect outliers in real time and display data to the user.

Robustness

To test the robustness of the software and the quality of the code, unit testing has been implemented with this software. The “test early, test often” approach allows new features to be tested against the functionality of the entire system very quickly and very easily [6]. A test suite developed with the aid of a test plan means full coverage of the testable methods in the system has been achieved. The test plan follows a format described by T. Hamilton in an article titled “Test Plan Template: Sample Document with Web Application Example [7]” and can be found in the appendix.

[1] B. Latte., S. Henning., M. Wojcieszak. (2019, Feb. 18) Clean Code: On the Use of Practices and Tools to Produce Maintainable Code for Long-Living. [Paper] *In: EMLS 2019: 6th Collaborative Workshop on Evolution and Maintenance of Long-Living Systems. , Stuttgart. Proceedings of the Workshops of the Software Engineering Conference 2019. ; pp. 96-99.* Available: <https://oceanrep.geomar.de/id/eprint/45829/>

[2] A. Baratloo., et al. (2015) “Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity.” *Emergency (Tehran, Iran) vol. 3,2: 48-49.* Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4614595/>

[3] Q. Nguyen., P. Eades., and S. Hong. (2012) "On the faithfulness of graph visualizations." *International Symposium on Graph Drawing. Springer, Berlin, Heidelberg*. Available: <https://link.springer.com/content/pdf/10.1007/978-3-642-36763-2_55.pdf>

[4] M. Bevan., J. Carter., and S. Harker. (2015) "ISO 9241-11 revised: What have we learnt about usability since 1998?" *International conference on human-computer interaction. Springer, Cham.* Available: <https://link.springer.com/chapter/10.1007/978-3-319-20901-2_13>

[5] D. Stone., et al. (2015) User interface design and evaluation. *Elsevier*. Available: [https://books.google.co.uk/books?hl=en&lr=&id=VvSoyqPBPbMC&oi=fnd&pg=PR21&dq=importance+of+user+interface&ots=d8M0SYsSR7&sig=i\_YqqRUztUrblzH0Q0hCp6pkVzE#v=onepage&q=importance%20of%20user%20interface&f=false](https://books.google.co.uk/books?hl=en&lr=&id=VvSoyqPBPbMC&oi=fnd&pg=PR21&dq=importance+of+user+interface&ots=d8M0SYsSR7&sig=i_YqqRUztUrblzH0Q0hCp6pkVzE%23v=onepage&q=importance%20of%20user%20interface&f=false)

[6] M. Olan. "Unit testing: test early, test often." *Journal of Computing Sciences in Colleges 19.2 (2003): 319-328.* Available: <https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=unit+testing&btnG=>

[7] T. Hamilton. (2022, Feb. 19) Test Plan Template: Sample Document with Web Application Example. *Guru99* [Online]. Available: <https://www.guru99.com/test-plan-for-project.html>

[8] A. Laurent., and M. Z. Hauschild. (2015) "Normalisation." *Life cycle impact assessment. Springer, Dordrecht, 2015. 271-300*. Available: <http://nozdr.ru/data/media/biblio/kolxoz/Cs/CsDb/Stephens%20R.%20Beginning%20database%20design%20solutions%20(Wiley,%202009)(ISBN%200470385499)(O)(552s)_CsDb_.pdf>

You should provide a precise description of the system developed. Note that this is likely to be different from those that you started with, since you now have considerably more knowledge and understanding. The final dissertation should therefore contain an updated set of requirements matching the final system delivered. You can list these as a *requirements definition* from the domain perspective but you should also derive a *specification* for the software. The software requirements specification establishes the basis for what the software product is to do (and is not expected to do). You should list any assumptions made about the problem and any system constraints. Overall your requirements, functional AND non-functional should be correct, complete, consistent, clear, feasible and objectively verifiable. Content depends on your project but could include:

* A complete set of function definitions (as use cases if preferred), as far as possible written so as to be testable
* Measurable and testable non-functional requirements
* Description of interfaces required such as with other software or systems
* Any specific user interface requirement
* User characteristics

The target to aim for here is to describe a solution that satisfactorily solves the problem. Ideally your solution will be convincing and creative. Your requirements could be the basis for a contract or handing to external developers to complete. The best dissertations will show outstanding work, approaching that of the best professionals.

* 1. Example sub heading

Use sub headings as appropriate

Subsub heading

For a sub sub heading just make it bold without numbering.

1. Design
   1. Design of the Detection Algorithm

The combined predictions of an ensemble of weak classifiers are used as the method to detect outliers in this experiment. These detectors generate a prediction for a piece of data (outlier or inlier), the combined predictions are combined to produce a final classification. Two methods of voting are used to produce the final classification. Experiments are performed to determine which method produces better score for accuracy, precision, recall and f1. The idea of concept drift must be considered since the VM cloud resource usage can vary throughout the day but also from day to day (due to factors such as new applications being deployed to these EC2 instances) [1].

There are 4 classification techniques in the Ensemble.

* **Moving Average**

This technique uses the average of the previous data points in the time series to classify the next. After the average is calculated, the standard deviation of the previous data points is calculated. The standard deviation is used as a threshold, if they next data point is less than or greater than the average calculated +/- the threshold then the data point is classified as an outlier [2]. The graph shows this technique in practice, the red lines represent the boundaries, and the red dots are the outliers detected.

Chart, line chart

Description automatically generated

Fig. 2 Moving Average outlier detection showing boundaries

* **Moving Median**

This technique follows the same steps as the previous except a median is calculated instead of the average [3]. The graph below shows this technique in practice, observe how the boundaries are like moving average but different outliers have been detected.

Chart, line chart

Description automatically generated with medium confidence

Fig. 3 Moving Median outlier detection showing boundaries

* **Moving Boxplot**

This technique takes several of the previous data points and generates a boxplot. The interquartile range is combined with the upper and lower quartiles to produce a threshold (1.5 \* the inter-quartile range). If the next data point is outside the threshold, then the data point is classified as an outlier [4].

1st Quartile

3rd Quartile

1.5 \* IQR (Upper bound)

1.5 \* IQR (Lower bound)

Fig. 4 Boxplot Outlier Detection Example

* **Histogram**

This technique plots histograms of subsets of the data. If a range in the histogram has a height less than a defined threshold, then the range is said to contain outliers [5]. If a range has a height below the threshold, but the ranges beside it have a height higher than the threshold then it is considered a borderline inlier.

**Chart

Description automatically generated**

Fig. 5 Histogram Based Outlier Detection

Ensemble Voting

For the detectors to work together they must vote on a classification for a specific data point. Two voting systems were proposed.

**1 – Majority Classification**

This voting technique labels outliers based on the classification made by the majority of detectors in the ensemble.

Equation: Majority Classification Voting Formula

Where ‘n’ is number of outliers in the ensemble and ‘t’ is the predefined number of detectors that must classify as an outlier.

Experiments performed using this technique proved this voting system to be efficient (running detection quickly) and accurate (producing a good score for accuracy) but did not generate good scores for recall, precision and f1 compare to voting system 2.

**2 – Combined Confidence**

Detectors in the ensemble run individually first, generating a prediction and a ‘confidence’ score. The confidence scores are combined to generate a final prediction. The formula behind this voting mechanism is described below.

Confidence is calculated by the distance between a data point and the threshold.

For predictions, -1(outlier) and 1(inlier), the above equation computes a minimum prediction of -n and a maximum of n. By visualising possible outputs on a spectrum, it can be said that an outlier score < 0 is likely to be an actual outlier.

0

N (max inlier)

-N (min outlier)

Experiments proved that this voting system performed the best, generating better scores for precision, recall and f1. Although it takes nearly double the time to complete the detection.

[1] A. Tsymbal. (2004). *The problem of concept drift: definitions and related work*. Computer Science Department, Trinity College Dublin 106.2: 58. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.9085&rep=rep1&type=pdf>

[2] Dr. Dataman. (2021, Apr, 18). *Anomaly Detection for Time Series. (1) Simple Moving Average.* [Online]. Available: [https://medium.com/dataman-in-ai/anomaly-detection-for-time-series-a87f8bc8d22e](file:///C:\Users\lreid\outlier-detection-in-virtual-machines\resources\%20https:\medium.com\dataman-in-ai\anomaly-detection-for-time-series-a87f8bc8d22e)

[3] Anomaly. (2016, Jan. 12). *Detecting Anomalies with Moving Median Decompsition*. [Online]. Available: <https://anomaly.io/anomaly-detection-moving-median-decomposition/index.html>

[4] A. Kliton, G. Shevlyakov and P. Smirnov. (2013). *Detection of outliers with boxplots*. 141-144. <https://www.researchgate.net/publication/261173084_Detection_of_outliers_with_boxplots>

[5] M. Goldstein, A. Dengel. (2012). *Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm*. Available: <https://www.goldiges.de/publications/HBOS-KI-2012.pdf>

* 1. Design of the Software

Database Design

|  |  |  |
| --- | --- | --- |
| **True Positive** | | |
| PK | true\_positive\_id | int |
| FK | detection\_ID | int |
|  | true\_positive\_datetime | datetime |

|  |  |  |
| --- | --- | --- |
| **Detection** | | |
| PK | detection\_ID | int |
|  | detector\_name | Varchar(50) |
|  | dataset\_name | Varchar(50) |
|  | true\_negative\_count | int |
|  | dataset\_size | int |

|  |  |  |
| --- | --- | --- |
| **False Negative** | | |
| PK | false\_negative\_id | int |
| FK | detection\_ID | int |
|  | false\_negative\_datetime | datetime |

|  |  |  |
| --- | --- | --- |
| **False Positive** | | |
| PK | false\_positive\_id | int |
| FK | detection\_ID | int |
|  | false\_positive\_datetime | datetime |

Fig. 3.2.1 Database Entity Relationship Diagram

Start

New detector/dataset selected?

No

End

Yes

Has detection been done before?

Get detection data from database

Yes

No

Calculate precision, recall, f1.

Do detection using specified detector / data

Output accuracy, precision, recall and f1 score.

Store detection data in database

Display detection on graph

Fig. 3.2.1 Flowchart for Experiments

Accuracy, precision, recall and F1 are calculated using the following equations.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

Equations used to evaluate the detector against the dataset.

Start

User terminated program?

Yes

End

No

New data?

No

Yes

Perform detection on data

Outlier?

Plot inlier on graph

Plot outlier on graph

Plot outlier on graph

Fig. 3.2.2 Real Time Outlier Detection Flowchart

UI Design

Detector Dropdown Box

Dataset Dropdown Box

Detection Results:

Accuracy

Precision

Recall

F1

Graph displaying outlier detection

Fig. 3.2.2 UI Design for Experimental/Testing Interface

Graphical user interface, text

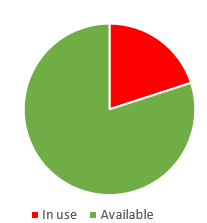
Description automatically generated

Fig. 3.2.3 Colours used to mark detection results on the graph

Dataset Dropdown Box

Detector Dropdown Box

CPU Usage



Graph displaying real time outlier detection

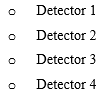
Fig 3.2.4 UI Design for Real Time Outlier Detection

Chart

Description automatically generated

Fig. 3.2.? UI Design for Real Time Detection

Radio buttons for selecting ensemble detectors



Dataset Dropdown Box

Detection Results:

Accuracy

Precision

Recall

F1

Graph displaying outlier detection

Fig 3.2.5 UI Design for Ensemble Testing Interface

This section should describe the design of your proposed system. Normally this several parts, depending on your project:

1. Architectural Description of the system – textual and/or diagrammatic. This could be a simple diagram showing the components and how they relate or it could describe the choice of architectural style or pattern used.
2. User Interface Design (if applicable). Show sketches of the design or screenshots with explanations of choices made, if necessary.
3. Software System Design. The role of each component and the interfaces between components should be described. There should be a clear correlation between your design and your specification.
4. Where applicable give a critical discussion of key design decisions/styles/patterns used; data model; UI design, external Interfaces, other important issues e.g. concurrency, event handling, data persistence, error and exception handling, fault tolerance, security, distribution of components.

The design should be linked to requirements and, where applicable give a critical discussion of key design decisions/styles/patterns used. There might be a data model, a UI design, details of external interfaces, and of other important issues e.g. concurrency, event handling, error and exception handling, security, data persistence. No particular notation or tool is mandated. A satisfactory design will show a grasp of the main design issues. For top marks aim for outstanding design documentation approaching that of the best professionals. Prove that you have a very strong grasp of the design issues and aim for documentation that could be passed on to a developer without the need for further explanation

1. Implementation

You should describe any languages, packages, and libraries etc. that are used in the development of your system. There is no need to describe your code in detail. You may highlight data types and implementation techniques that are of special interest. If appropriate, you may provide:

* 1. Development Methodology

The software solution has been developed using elements of the ‘Agile’ development methodology. ‘Incremental’ and ‘Iterative’ development practices were used to gradually implement and test features of the application. This allowed an initial working prototype with minimal features to be gradually developed into the final solution. Features implemented in older versions could be revisited and improved [1].

* 1. Implementation Language

Python was selected as the implementation language for this software solution because of its versatility. It is commonly used for web-development, data analysis and machine learning. Python “has become a staple in data science”, with many tools making data visualisation quick and effortless [2]. It is used in web development to implement back end functionality and libraries exist to allow secure communications with databases and software. Additionally, it is the most popular computer programming language (as of April 2022) [3], there is a large community of Python developers that constantly generate and update resources and documentation.

* 1. Development Environment

Visual Studio Code was selected as the development environment. It has features such as automatic code completion, bracket matching, auto-indentation and an interactive debugger. Visual Studio Code has support for Git, source control is possible without leaving the IDE and changes from previous commits are highlighted, making it easy for the developer to modify changes [4]. Additionally, according to a stackoverflow survey, Visual Studio Code is the most popular IDE among professional developers [5].

* 1. Software Libraries

To aid the implementation of this solution, a number of software libraries were imported.

**Pandas**

This software library is used for data analysis and manipulation [6]. The ‘Dataframe’ objects created using this library played a fundamental role in this software implementation. Dataframes are especially useful for temporarily storing data coordinates and outlier detection results. As such, almost all data flowing through the system is in a pandas Dataframe object.

**Plotly**

This library is used for generating interactive graphs [7]. The software solution is a monitoring tool thus graphs are an important component in aiding with data readability. Line plots, scatter plots and pie charts are implemented throughout the software solution using this library. Additionally, interactive live update graphs are possibly with Plotly, this feature was used to display the real time detection of outliers.

**Dash**

Interactive dashboard applications can be created with the Dash library. The front end can be implemented by defining the ‘layout’ of the web-app using “Dash HTML components”. Back end functionality is implemented using “Callback” functions [8]. This library was used to implement the front end (user interface) and the ‘Callback’ functions were used to bridge the front end to the data generated by the detection algorithms.

**Sqlite3**

The software solution required storage for the detection results so that complex, resource intensive processes did not need to run redundantly to generate pre-existing data. The sqlite3 software library is used to implement a server-less, self-contained database [9]. It is “small, fast, fully featured, SQL database” which serves as a perfect solution for the software’s storage needs.

**Pycaret**

The Pycaret Anomaly Detection module was implemented early in the development of this project. It contains an array of unsupervised traditional detection techniques such as SVM and KNN. The detection techniques can be applied to a dataset and outliers are predicted [10]. The techniques in this library were used to build the simple graphing and scoring mechanisms behind the prototype of this application, they were also used a benchmark to test the ensemble technique proposed in this dissertation.

**Sklearn**

This library is used for machine learning in python [11]. In the experimental space of the software, there is a section for supervised learning. The supervised learning methods in this section were implemented using this library.

**Requests**

To simulate a real time stream of data read in from an external location, a server has been implemented that handles GET requests and returns CPU usage. This is accessed by the software using the ‘Requests’ library. ‘Requests’ allows HTTP requests to be sent and the return data can be stored as a variable and used in the application [12].

* 1. Architecture

The development of software involves collaboration among teams of developers and testers. Code bases can become increasingly difficult to understand if pre-defined conventions are not followed [13]. Hence, clean architecture has been used in the implementation of the software solution.

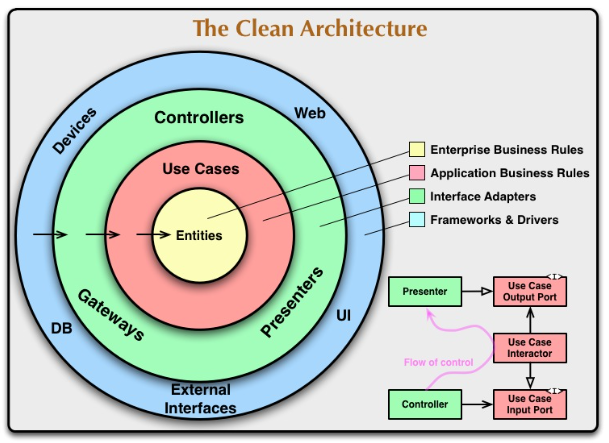


Fig. 4.6.1. The Clean Architecture [14].

Fig. 4.6.1 Describes clean architecture in a diagrammatic form. It consists of a group of labelled concentric circles that represent different components of a software system.

The circle at the centre of the diagram is labelled ‘Entities’. At this level, classes are implemented that do not depend on any other component of the system. The detection algorithms have been implemented at this level.

The next circle is labelled ‘use-cases’. At this level, classes are implemented that orchestrate the flow of data to and from the entities in the inner most layer. The classes at this layer act as an interface between data and detection.

The next circle is labelled ‘Interface Adapters’. Software at this level is used to convert data to different formats to and from the ‘use-cases’ classes. Methods to plot data on graphs, or pass input to use cases are implemented at this level.

The outermost circle is ‘Frameworks and Drivers’. At this level, classes concerning the web application are implemented.

Clean architecture is achieved by applying ‘The Dependency Rule’ to the structure of the application. This rule states that “source code dependencies can only point inwards” [14]. Therefore, the classes implemented with the software solution fall into a hierarchical pattern. This can be observed as a ‘cone’ of the clean architecture diagram described in Fig. 4.6.1.

Diagram, timeline

Description automatically generated

Fig. 4.2.2. Architecture Diagram of Implemented Classes.

Fig. 4.2.2 shows the architecture of the entire software solution (areas are shaded to match those labelled in Fig. 4.2.1). It follows ‘The Dependency Rule’ described as part of ‘Clean Architecture’. Classes in the lower levels do not depend on classes higher up. With the classes implemented this way, the software is easily maintained, and new features are easily implemented.

* 1. Description of Component Implementation

[1] J. Shore., and S. Warden. (2021) The art of agile development. "O'Reilly Media, Inc." Available: <https://books.google.com/books?hl=en&lr=&id=i3ZIEAAAQBAJ&oi=fnd&pg=PT4&dq=agile+development&ots=VCWTBZ_I4Z&sig=ruhYeT7IqwxD1_HXacwqBCLYvoE>

[2] Coursera. (2022, Mar. 21) What Is Python Used For? A Beginner’s Guide [Online]. Available: <https://www.coursera.org/articles/what-is-python-used-for-a-beginners-guide-to-using-python>

[3] Tiobe. (2022, April) TIOBE Index for April 2022 [Online]. Available: <https://www.tiobe.com/tiobe-index/>

[4] Visual Studio Code. (n.d.) Why did we build Visual Studio Code? [Online]. Available: <https://code.visualstudio.com/docs/editor/whyvscode>

[5] Insights – Stackoverflow. (2021) Stack Overflow Developer Survey 2021 – Most Popular Technologies. [Online]. Available: <https://insights.stackoverflow.com/survey/2021#technology>

[6] Pandas. (2022) Pandas Documentation. [Online]. Available: <https://pandas.pydata.org/docs/>

[7] Plotly. (n.d.) Plotly Python Open Source Graphing Library. [Online]. Available: <https://plotly.com/python/>

[8] A. Tomar. (2021, Mar. 17) Dash for Beginners: Create Interactive Python Dashboards. [Online]. Available: <https://towardsdatascience.com/dash-for-beginners-create-interactive-python-dashboards-338bfcb6ffa4>

[9] SQLite. (n.d.) What Is SQLite? [Online]. Available: <https://www.sqlite.org/index.html>

[10] Pycaret. (n.d.) Anomaly Detection *– Pycaret Quickstart* [Online]. Available: <https://pycaret.gitbook.io/docs/get-started/quickstart#anomaly-detection>

[11] Scikit-learn (n.d.) Scikit-learn – Getting Started [Online]. Available: <https://scikit-learn.org/stable/getting_started.html>

[12] Python-Requests. (n.d.) Requests: HTTP for Humans. [Online]. Available: <https://docs.python-requests.org/en/latest/>

[13] P. Ivanics. (n.d.) An Introduction to Clean Code Architecture. *Department of Computer Science – University of Helsinki* [Online]. Available: <http://pivanics.users.cs.helsinki.fi/portfolio/docs/publications/Peter_Ivanics-Clean_Software_Architecture.pdf>

[14] R. C. Martin. (2012, Aug. 13) The Clean Architecture. *The Clean Code Blog.* [Online]. Available: <https://blog.cleancoder.com/uncle-bob/2012/08/13/the-clean-architecture.html>

Choice of implementation language(s)/ development environment(s)

1. Use of software libraries;
2. Key implementation decisions
3. A description of how some important functions and algorithms were implemented.
4. A description of how each component is implemented.

Program code can be accessed by the assessors via the git repository **so there is no need to include code listings**. It is recommended that you comment code appropriately (not excessively). Programs should be written in a clear style with good program structure and well-defined data structures. The program code should reflect its design and show an understanding of relevant implementation issues.

1. Testing
   1. Continuous Integration – Automated Test Suite

Continuous integration was paired the ‘Agile’ development lifecycle, this allowed changes to be tested against the entire repository within minutes. New features being implemented where committed to the main branch of the repository (often multiple times a day) for fast feedback on code changes. This was made possible with the implementation of an automated test suite [2].

The implementation of the test suite came in the form of unit tests written in python. White-box testing techniques were used to design the automated test suite with the goal of achieving high code coverage. Unit testing is a method of individually testing components of software to validate code acts as expected, unit tests are designed to run quickly, thus new code changes can be validated against the code base within seconds [3].

Before Continuous Integration was implemented with this project, a testing plan was developed. This plan contains the methods/techniques considered when designing and implementing the unit tests.

* 1. Test Plan for Design and Implementation of the Automated Test Suite

Introduction

The test plan follows white-box testing principles. Test case designs are based on the flow of code of the methods implemented. Test adequacy will be determined by the portion of the code that is exercised. For newly implemented features, black-box testing techniques are used to derive initial test cases. Then, an iterative approach is used to improve code coverage by analysing lines and branches missed by the initial test cases. This testing approach verifies functionality exists and works as intended.

Scope

To achieve good code coverage, unit tests are written for all methods of a feature implemented. In some instances, unit tests may be inefficient for some of the implemented features (i.e. graph plotting functions). These methods are not included in the automated test suite and are tested a different way.

Quality Objective

The ultimate objective of the automated test suite is to ensure methods act as expected and do not break any functionality in the rest of the code base.

Good coverage will guarantee the code is robust and acts as intended. Therefore, a threshold of 90% code coverage over the entire application will be used to verify the code base is well exercised in automated tests.

For maintainability purposes, conventions are used for test code architecture.

* Test code is modularised with respect to the implemented code (unit tests are grouped and ordered).
* Test scripts are named by prefixing ‘test\_’ to the name of the class being tested.
* The unit test method name refers to the purpose of what the test is attempting to achieve/assert.

Conventions are included as part of the quality objective because it makes the test code maintainable. Maintainability is important when writing code in any form since other developers can review, edit or understand failures with a given test case without having to dissect code line by line.

Test Methodology

White box testing techniques are used to enhance the automated test suite, but initial commits of features include tests written with black box testing techniques. The requirements of a method are used as the test oracle for the initially implemented test cases.

* Black box testing techniques
  + Boundary Value Analysis

This black box testing technique ensures there are no defects with a method when extreme or boundary values are used as input.

* + Equivalence-Class Partitioning

This method of testing aims to use test code to produce a full set of function outputs with the minimum amount of test cases.

* White box testing techniques
  + Control-Flow Coverage

Specifically ‘Path Coverage’ is used to design test cases. Path coverage ensures all nodes in a method are traversed i.e. every possible condition for a conditional branch is passed to the function.

‘Coverage.py’ - <https://coverage.readthedocs.io/en/6.3.2/>, is a python test coverage tool. Coverage.py will be used to generate reports on code coverage and determine test adequacy.

Test Deliverables

There are 2 phases involved with the unit test implementation.

* During feature implementation
  + Black box testing techniques are used to implement an initial set of test cases.
* After feature implementation
  + Coverage.py report is generated to determine code coverage using the black box testing techniques.
  + If coverage < 90%, white box testing techniques are applied to improve test adequacy.
    - Coverage report generated again. Iterate until coverage >= 90%.

At each phase of implementation, testing is performed and coverage can be determined.

* 1. Test Environment – Docker

The automated test suite runs in a Docker container, the test suite is triggered when a commit is made to the Gitlab repository. Fig. 5.3.1 shows this pipeline and fig 5.3.2 shows the ‘Gitlab-CI’ configuration. Personal hardware is used to host the Docker container. Without the budget constraint of this project, the repository would be configured to spin up a Docker instance in the cloud, run the tests, and then tear it down. This would remove the need for personal hardware being used as the runner.

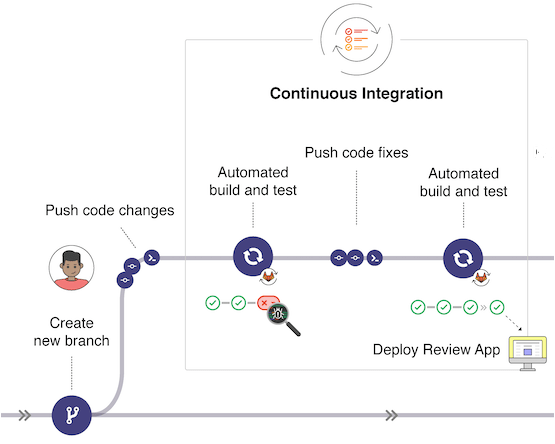


Fig. 5.3.1 Continuous Integration Pipeline [1]

[1] Gitlab Docs. (n.d.) CI/CD Concepts [Online]. Available: <https://docs.gitlab.com/ee/ci/introduction/index.html#continuous-integration>

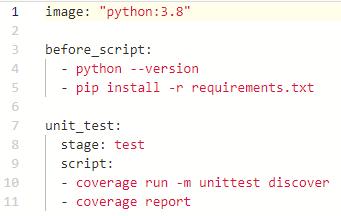


Fig. 5.3.2 Gitlab-CI Configuration.

* 1. Automated Test Suite Result

The output of the automated test suite executed on the Docker container is shown in Fig. 5.4.1. A large number of unit tests run in a very short amount of time. The coverage report shows good test adequacy (>=90% total) over the code so the quality objective of the test plan has been achieved.

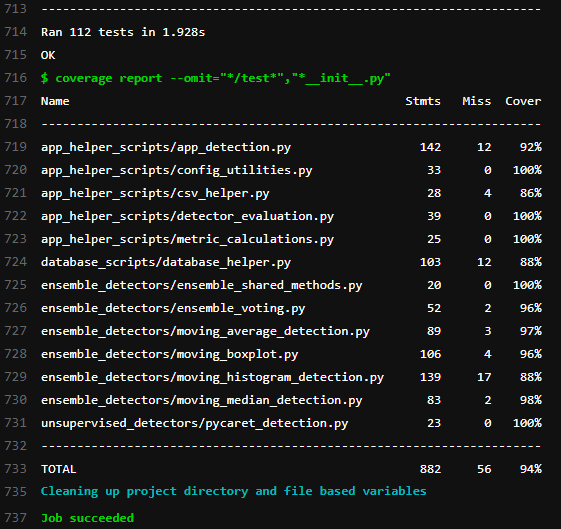


Fig. 5.4.1 Output of Automated Test Suite Execution.

SCREENSHOT OF TEST EXECUTION ON CI LOG

[2] S. M. Mohammad. (2016, Jul. 3) Continuous Integration and Automation. International Journal of Creative Research Thoughts *(IJCRT), ISSN:2320-2882, Volume.4, Issue 3, pp.938-945.* Available: <https://ssrn.com/abstract=3655567>

[3] D. Sale. (2014) Testing Python: Applying Unit Testing, TDD, BDD and Acceptance Testing. *John Wiley & Sons*. Available: <https://books.google.com/books?hl=en&lr=&id=vtj1AwAAQBAJ&oi=fnd&pg=PT14&dq=unit+testing+python&ots=atGgN0KHm8&sig=HpYOAgy20seK3NMnO4DbPqzIDFw>

This section will be judged in tandem with other evidence including evidence of unit tests and/or test documentation on the Repo. There should be a discussion of Test Approach e.g. unit testing, system testing, regression testing etc; Test cases should be described and justified; Include Testing tools used and provide evidence that testing coverage was complete. Provide proof that testing was completed, either showing sample test history and/or describing automated tests.

1. System Evaluation - Experimental Results

To evaluate the performance of the implemented ensemble detection technique, and to determine which voting system works best, two experiments were conducted. One was on labelled datasets and one on unlabelled datasets.

* 1. Labelled Data - CPU Utilization Outlier Detection Experiment

## Abstract

This experiment tests the effectiveness of a newly implemented technique of detecting outliers. This technique involves detecting outliers using an ensemble of ‘weak’ classifiers that work together and vote on whether a data point is an outlier or an inlier. The experiment is performed on Cloud platform CPU usage. Techniques are evaluated using accuracy, recall, precision and f1. This experiment determines which voting system works best out of two implemented. Results show that the ensemble detection method implemented can detect outliers in the CPU usage data. It sometimes produces good scores but does not perform well against unstable data. This experiment determines that a ‘Combined Confidence’ voting system produces the best scores.

## Introduction to the Experiment

**Background Information**

The data being analyzed is Amazon Elastic Compute Cloud (EC2) CPU usage. EC2 is a service provided by Amazon used for on-demand cloud infrastructure. Customers use the platform for its compute power, running multiple different kinds of operating systems for various applications [1].

Amazon Web Services (AWS) provides many tools for analyzing the metrics of an EC2 instance. CPU Utilization is arguably the most important metric, since it “identifies the processing power required run an application on a selected instance [2]”. Problems with an EC2 instance, or an application running on one, can often be identified by a discrepancy in CPU usage [3].

This CPU utilization data is labelled, meaning the outlier detection technique applied to the data can be evaluated based on accuracy, precision, recall and f1 score [4].

**How is the experiment carried out?**

This experiment is carried using the application developed as part of this project. The purpose of this application is to apply outlier detection algorithms to datasets defined by the user.

Graphical user interface

Description automatically generated

Fig. 1 Screenshot showing application used to apply outlier detection.

## Hypothesis

The implemented outlier detection method (ensemble) is an effective outlier detector for labelled datasets and will generate good scores for accuracy, precision, recall and f1.

The proposed scoring method for the ensemble of detectors will generate better scores than the previously implemented solution (Combined confidence will outperform majority classification).

## Methods

The methods used to perform detection are described in section 3.1.

**Method of Scoring the Detector**

The datasets being analyzed have been labelled by the Numenta Anomaly Benchmark (NAB). NAB is a platform for testing detection techniques on time series data. It contains 58 labelled datasets used for scoring algorithms [5]. Labelling data is expensive [6], but these datasets are available for free from NAB and are crucial to this experiment.

The software used to perform the experiments uses the labels provided by NAB to plot the outliers. Fig. 7 shows a sample of time series data with classifications. The key for this data is represented by Fig. 6.

Graphical user interface, text

Description automatically generated

Fig. 6 Outlier detection classification key

Graphical user interface, chart, line chart

Description automatically generated with medium confidence

Fig. 7 Graph demonstrating how the application represents the detection data.

## Evaluating Results

The software calculates accuracy, precision, recall and f1 score to determine the overall performance of the detector. The results of these calculations and graphs generated by the software can be found in Appendix A.

**Evaluating the Voting Methods**

The ‘Combined Confidence’ voting system produces better scores than the ‘Majority Classification’ voting system concluding that the former is a better solution, as hypothesized. Tables 1 and 2 in Appendix A provide detailed scores for each dataset and table 4 provides a side-by-side comparison. ‘Combined Confidence’ produces better scores for precision, recall and f1. Table 3 shows that the ‘Majority Classification’ system cannot outperform moving average (one of the detectors in the ensemble) proving this method is ineffective. ‘Majority Classification’ has a higher average accuracy, but false negatives are crucial when analyzing CPU usage [7], therefore this metric is less useful than precision, recall and f1.

‘Majority Classification’ is more efficient than ‘Combined Confidence’, the simplicity of its voting system results in a lower execution time (less than half the time taken compared to the confidence technique). Some improvements may be required to improve the detection time so that this technique can work for real time outlier detection.

**Evaluating the Effectiveness of the Ensemble**

The ensemble technique of detecting outliers is sometimes very effective. In Fig. 7, graphs I, V, VIII and IX show that the detection is working and good scores for recall, precision and f1 are generated.

This technique is sometimes ineffective, especially against unstable datasets. In Fig. 7, graphs II, III and VII show that the detection has failed, and the ensemble of detectors are ineffective. Although the ensemble produces weak scores in these datasets, table 3 shows that the ‘moving histogram’ detector produces good scores. An improvement to the voting system, by potentially adding weighted confidences, could produce better detection in these graphs.

Observations of Fig. 7 show that the detector is very nearly producing perfect scores for some datasets. Graph VI shows that a false positive detection was made 1 data point in the time series away from the false negative. This would have produced a perfect score for this dataset. Similarly, in graphs VIII and IX, the detectors would have produced perfect scores if they had correctly classified the second false negatives.

## Conclusion

This experiment evaluated the newly implemented Ensemble technique of detecting outliers and determined that it is effective in detecting outliers and producing good scores in some datasets.

It was found that this technique is ineffective in detecting outliers in unstable datasets, but some detectors within the ensemble are more effective in unstable datasets than others. Meaning improvements to the voting system could improve upon this issue. In other datasets, perfect scores are almost achieved.

Results show that ‘Combined Confidence’ produces better scores than ‘Majority Classification’ as hypothesized. The intricacy of the voting system means that it takes longer to perform detection, but some optimization techniques could improve this score.

[1] AWS Amazon. (2022). *Amazon EC2 – Secure and resizable compute capacity for virtually any workload* [Online]. Available: <https://aws.amazon.com/ec2/>

[2] AWS Amazon (2022). *Monitor Amazon EC2* [Online]. Available: <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/monitoring_ec2.html>

[3] Ionos (2020, Feb. 24) *High CPU usage: What does this mean?* [Online]. Available: <https://www.ionos.com/digitalguide/server/know-how/cpu-usage/>

[4] Y. Sasaki. (2007). The truth of the F-measure. Teach Tutor Mater. Available: <https://www.researchgate.net/publication/268185911_The_truth_of_the_F-measure>

[5] Numenta. (2015). *The Numenta Anomaly Benchmark – White Paper.* 1. Available: [https://numenta.com/assets/pdf/numenta-anomaly-benchmark/NAB-Business-Paper.pdf#:~:text=The%20Numenta%20Anomaly%20Benchmark%20%28NAB%29%20is%20an%20open,NAB%3A%20the%20labeled%20dataset%20and%20the%20scoring%20system.](https://numenta.com/assets/pdf/numenta-anomaly-benchmark/NAB-Business-Paper.pdf%23:~:text=The%20Numenta%20Anomaly%20Benchmark%20%28NAB%29%20is%20an%20open,NAB%3A%20the%20labeled%20dataset%20and%20the%20scoring%20system.)

[6] Cloudfactory. (n.d.). *The Ultimate Guide to Data Labelling for Machine Learning* [Online] Available: <https://www.cloudfactory.com/data-labeling-guide>

[7] P. Huilgol. (2019, Aug. 24) *Accuracy vs. F1-Score* [Online]. Available: <https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2>

* 1. Unlabeled Data - Dengue Fever Rate Experiment

## Abstract

The aim of this experiment is to test the effectiveness of a newly implemented technique of detecting outliers. This technique involves detecting outliers using an ensemble of ‘weak’ classifiers that work together and vote on whether a data point is an outlier or an inlier. This technique is tested on Dengue Fever rates in regions of Vietnam. Observations of generated graphs show that this technique is effective in detecting outliers. Obvious outliers and some subtle outliers can be detected using this method but on rare occasions an outlier is missed and there are many false alarms. A comparison with a traditional classifier (KNN) proves that this method of detecting outliers is of good standard.

## Introduction

**Background Information**

This document contains details of newly implemented outlier detection techniques on Dengue fever rates in the Bac Lieu and An Giang regions of Vietnam.

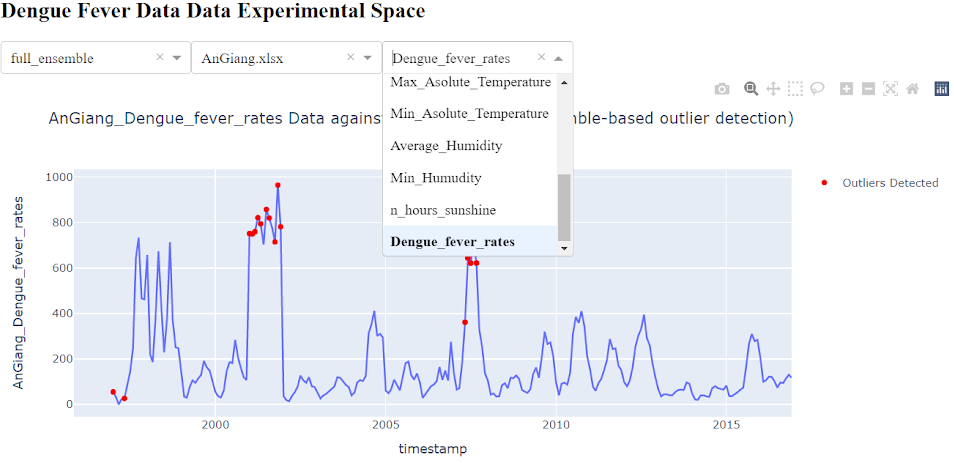
The Dengue Fever is a viral disease that is carried by mosquitos. It is widespread throughout tropical regions. Local environmental factors have an impact of the spread and severity of this virus and it is leading cause of hospitalisation and deaths in the areas that it affects [1].

The data consists of details about the climate of these regions over a 20-year period. Data such as average temperature, rainfall, humidity and the fever rate itself will be considered in this experiment. The data is unlabelled. The techniques used will be unsupervised and will not require any training. The ultimate goal of this experiment is to run outlier detection on these datasets and see if the outliers correlate between different datasets, if they do, then the outlier detection is working.

**How is the experiment carried out?**

This experiment is carried using the application developed as part of this project. The purpose of this application is to apply outlier detection algorithms to datasets defined by the user.

The application plots the data and generate scores for the chosen method. A score is not generated for the detection methods on these datasets since they are unlabelled.



Screenshot of application used to perform detection and generate graphs.

The outlier detection methods in the ensemble are implemented using python. These methods work individually first to make a prediction with a confidence score. A voting system, also implemented using python, determines the final classification.

## Hypothesis

The implemented outlier detection method (ensemble) is an effective outlier detector for unlabelled datasets.

## Methods

The methods used to perform detection are described in section 3.1.

## Evaluation of An Giang Data

Graphs showing results generated by the software can be found in Appendix B Fig. 1.

The graphs above show that outliers have been detected in peaks and troughs throughout the time series, indicating that (possible) actual outliers have been detected. Most notably, a cluster of detections were made in graph I in the spike between 2000 and 2005. This correlates with the outlier detected in graph III indicating rainfall caused a spike in the fever rate. More notably, outliers were indicated in graph IV between 2010 and 2015 where there seems to be no raining days. In graph III, a spike in rainfall is detected, which correlates with the spike in evaporation detected in graph VI around the same time.

A great number of detections have been made in graph V compared with the rest, the detection technique may be ineffective against such unstable data. But a cluster of detections are shown in the 2005 to 2010 period. There is a trough in the time series which correlates with the detection made in graph II where an outlier is detected in average humidity. The outlier detected here does not look irregular when plotted since it is not major peak or a trough but could be expected to be an actual outlier because of the irregular temperature.

## Evaluation of Bac Lieu Data

Graphs showing results generated by the software can be found in Appendix B Fig. 2.

The graphs show that more outliers are detected here compared with the An Giang data, especially in graph V. Besides this, the outliers detected are in the peaks and troughs of the data. Detections were made around the spikes in fever rates in graph I, but the detector is failing to correctly classify the top of some peaks.

Similarly to graph V in the An Giang Data, the detector has marked many data points as outliers, solidifying the fact that this detector may be ineffective against unstable data. Besides graph V there are a number of correlations between the graphs. The initial spike in dengue fever rates in graph I correlate with the average humidity in graph II and the drop in no. raining days in graph IV. These correlations are marked as outliers by the detectors. Again, the trough in graph IV is picked up by the detector as well as a sudden spike in total rainfall in graph III.

## Comparison with a Traditional Classification Technique – KNN

To test the effectiveness of this newly implemented ensemble, KNN outlier detection has been applied to the datasets [2].

## Comparison of Results

KNN failed to detect some of the obvious outliers (peaks/troughs) that are detected by the ensemble method. In graph IV of the An Giang data, the ensemble method correctly identifies outliers in the 3-4 year stretch where there were no raining days, but KNN fails to detect this data as anomalous. Similarly, KNN detects two spikes in dengue fever rates in the An Giang region and misses a major peak in around 2007, the ensemble technique detects this spike.

KNN performs better for graph V in both regions. The ensemble method detects a large number of outliers whereas KNN detects a few in areas of the graphs that (appear to be) actual outliers. KNN appears to detect the top of peaks better than the ensemble method.

Something important to note when comparing these results is that the ensemble method takes much less time to run than KNN. It took KNN ~6 seconds to process each dataset and it took the ensemble ~1 second using the same hardware.

## Conclusion

The ensemble method of detecting outliers is effective in detecting outliers in unlabelled datasets. It is difficult to say to what extent it is effective since accuracy, precision, recall and f1 are impossible to calculate without labels. But, by comparing this method with a traditional classifier, and observing peaks and troughs within the datasets, it can be said that the ensemble is detecting outliers in the correct places. For most datasets the ensemble appears to be performing the same or even better than the traditional KNN detector but it is clear that the ensemble needs more work around unstable datasets. Another important thing to note is that with similar results, the ensemble performs detection up to 6x faster than KNN.

## References

[1] World Health Organization. (2022, Jan. 10). Dengue and Severe Dengue [Online]. Available: <https://www.who.int/en/news-room/fact-sheets/detail/dengue-and-severe-dengue>

[2] Harrison, O. (2018, Sep. 10). Machine Learning Basics with the K-Nearest Neighbours Algorithm. [Online]. Available: <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>

## Appendix A – EC2 CPU Experiment Results

|  |  |
| --- | --- |
| Chart, histogram  Description automatically generated | Chart  Description automatically generated |
| I. Detection result for Numenta VM1 | II. Detection result for Numenta VM2 |
| A picture containing text, sky, bunch, line  Description automatically generated | Chart, bar chart, histogram  Description automatically generated |
| III. Detection result for Numenta VM3 | IV. Detection result for Numenta VM4 |
| Graphical user interface, chart  Description automatically generated | Timeline  Description automatically generated with medium confidence |
| V. Detection result for Numenta VM5 | VI. Detection result for Numenta VM6 |
| Chart  Description automatically generated | Graphical user interface  Description automatically generated with low confidence |
| VII. Detection result for Numenta VM7 | VIII. Detection result for Numenta VM8 |

Chart

Description automatically generated

IX. Detection result for Numenta VM9

Fig. 8 Graphs showing CPU utilization over time with outliers detecting using the Ensemble and ‘Combined Confidence’ voting.

Table 1 Results for ensemble detection with ‘Combined Confidence’ voting.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| VM NAME | Accuracy | Recall | Precision | f1 | Time to execute |
| Numenta VM1 | 99.97 | 50.00 | 50.00 | 50.00 | 14.3014 |
| Numenta VM2 | 66.10 | 100.00 | 0.15 | 0.29 | 21.6323 |
| Numenta VM3 | 43.23 | 100.00 | 0.09 | 0.17 | 25.6607 |
| Numenta VM4 | 99.20 | 0.00 | 0.00 | 0.00 | 14.0828 |
| Numenta VM5 | 99.90 | 50.00 | 20.00 | 28.57 | 14.4061 |
| Numenta VM6 | 99.98 | 0.00 | 0.00 | 0.00 | 14.2985 |
| Numenta VM7 | 70.26 | 0.00 | 0.00 | 0.00 | 18.5497 |
| Numenta VM8 | 100.00 | 50.00 | 100.00 | 66.67 | 13.734 |
| Numenta VM9 | 99.98 | 50.00 | 50.00 | 50.00 | 13.8773 |
| Average | 86.51 | 50.00 | 27.53 | 24.46 | 16.7270 |

\* Numenta VM7 is excluded from the average since there are no true positives and an f1 score cannot be calculated \*

Table 2 Results for ensemble detection with ‘Majority Classification’

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| VM NAME | Accuracy | Recall | Precision | f1 | Time to execute |
| Numenta VM1 | 99.98 | 50.00 | 50.00 | 50.00 | 5.3486 |
| Numenta VM2 | 77.75 | 100.00 | 0.22 | 0.44 | 7.2271 |
| Numenta VM3 | 78.17 | 100.00 | 0.23 | 0.45 | 8.7147 |
| Numenta VM4 | 99.93 | 0.00 | 0.00 | 0.00 | 5.4163 |
| Numenta VM5 | 99.88 | 50.00 | 16.67 | 25.00 | 5.7832 |
| Numenta VM6 | 100.00 | 0.00 | 0.00 | 0.00 | 4.8424 |
| Numenta VM7 | 76.81 | 0.00 | 0.00 | 0.00 | 7.2695 |
| Numenta VM8 | 100.00 | 50.00 | 100.00 | 66.67 | 5.0816 |
| Numenta VM9 | 99.85 | 50.00 | 14.29 | 22.22 | 6.8909 |
| Average | 92.49 | 44.44 | 20.16 | 18.31 | 6.2860 |

\* Numenta VM7 is excluded from the average since there are no true positives and an f1 score cannot be calculated \*

Table 3 F1 scores (%) of Individual Detectors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| VM Name | Average | Median | Boxplot | Histogram |
| Numenta VM1 | 50.0 | 50.0 | 0.0 | 0.0 |
| Numenta VM2 | 0.1 | 0.2 | 0.5 | 66.7 |
| Numenta VM3 | 0.1 | 0.1 | 0.0 | 0.0 |
| Numenta VM4 | 0.0 | 0.0 | 0.3 | 0.0 |
| Numenta VM5 | 25 | 25.0 | 8.2 | 0.0 |
| Numenta VM6 | 0.0 | 0.0 | 4.3 | 0.0 |
| Numenta VM7 | 0.0 | 0.0 | 0.0 | 100.0 |
| Numenta VM8 | 66.7 | 50.0 | 1.8 | 0.0 |
| Numenta VM9 | 50.0 | 50.0 | 9.1 | 0.0 |
| Average | 24.0 | 22.0 | 3.0 | 8.3 |

Table 4 Comparison of Voting Systems

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Voting System | Accuracy | Recall | Precision | f1 |
| Combined Confidence | 86.51 | 50.00 | 27.53 | 24.46 |
| Majority Classification | 92.49 | 44.44 | 20.16 | 18.31 |

Table 5 VM Data Naming Convention

|  |  |  |
| --- | --- | --- |
| S/No | VM ID | VM NAME |
| 1 | ec2\_cpu\_utilization\_5f5533 | Numenta VM1 |
| 2 | ec2\_cpu\_utilization\_24ae8d | Numenta VM2 |
| 3 | ec2\_cpu\_utilization\_53ea38 | Numenta VM3 |
| 4 | ec2\_cpu\_utilization\_77c1ca | Numenta VM4 |
| 5 | ec2\_cpu\_utilization\_825cc2 | Numenta VM5 |
| 6 | ec2\_cpu\_utilization\_ac20cd | Numenta VM6 |
| 7 | ec2\_cpu\_utilization\_c6585a | Numenta VM7 |
| 8 | rds\_cpu\_utilization\_cc0c53 | Numenta VM8 |
| 9 | rds\_cpu\_utilization\_e47b3b | Numenta VM9 |

## Appendix B – Dengue Fever Rates Experiment

|  |  |
| --- | --- |
| Graphical user interface, chart  Description automatically generated | Graphical user interface  Description automatically generated |
| I. Detection result for An Giang Dengue Fever Rate | II. Detection result for An Giang Average Humidity |
|  |  |
| Graphical user interface  Description automatically generated | Chart  Description automatically generated |
| III. Detection result for An Giang Total Rainfall | IV. Detection result for An Giang No. Raining Day |
|  |  |
| A picture containing different, various, colorful, several  Description automatically generated | Graphical user interface, application  Description automatically generated |
| V. Detection result for An Giang Average Temperature | VI. Detection result for An Giang Total Evaporation |

Fig. 1 Ensemble Detection on An Giang Dengue Fever Data

|  |  |
| --- | --- |
|  |  |
| I. Detection result for Bac Lieu Dengue Fever Rate | II. Detection result for Bac Lieu Average Humidity |
|  |  |
|  |  |
| III. Detection result for Bac Lieu Total Rainfall | IV. Detection result for Bac Lieu No. Raining Day |
|  |  |
|  |  |
| V. Detection result for Bac Lieu Average Temperature | VI. Detection result for Bac Lieu Total Evaporation |

Fig. 2 Ensemble Detection on Bac Lieu Dengue Fever Data

|  |  |
| --- | --- |
|  |  |
| I. Detection result for An Giang Dengue Fever Rate | II. Detection result for An Giang Average Humidity |
|  |  |
|  |  |
| III. Detection result for An Giang Total Rainfall | IV. Detection result for An Giang No. Raining Day |
|  |  |
|  |  |
| V. Detection result for An Giang Average Temperature | VI. Detection result for An Giang Total Evaporation |

Fig. 3 KNN Outlier Detection on An Giang Dengue Fever Data

|  |  |
| --- | --- |
|  |  |
| I. Detection result for Bac Lieu Dengue Fever Rate | II. Detection result for Bac Lieu Average Humidity |
|  |  |
|  |  |
| III. Detection result for Bac Lieu Total Rainfall | IV. Detection result for Bac Lieu No. Raining Day |
|  |  |
|  |  |
| V. Detection result for Bac Lieu Average Temperature | VI. Detection result for Bac Lieu Total Evaporation |

Fig. 4 KNN Outlier Detection on Bac Lieu Dengue Fever Data

Provide a summary evaluation of the success of the project with respect to criteria identified in the introduction. Different projects will have a different emphasis. In all cases you are expected to provide empirical results and to draw conclusions from those results. You may use your software to generate experimental results. Be sure to describe the methodology of your evaluation or experimentation. An experiment is typically described in terms of its goals, the hypotheses being tested, the subject of the experiment, what is being measured and what is controlled, the results obtained and the analysis and interpretation of those results. A discussion of the significance of your experimental results may be appropriate or why the new system you have developed improves on what was already there. Do your results agree with other previous work or ideas? How does your system compare with similar ones?

Alternatively (or additionally), you can assess the product in terms of how it compares with other similar products and/or in terms of user feedback (e.g. via a survey or interviews) or some measurable quality aspect such performance efficiency or reliability.

Draw conclusions on the *process* used in the project as well. What went well? What did not go well? What are the strengths of your solution or conclusions? What are the weaknesses? Suggestions for further work should also be discussed. You can be critical and draw a negative conclusion. Not all projects will be successful. A well-explained failure is as an acceptable an outcome as a spectacular success. Assessors are looking for excellence in a critical appraisal of the work and a convincing argument for the significance of contribution in the context of wider work. This section should be objective, fair and comprehensive.

In all cases, societal implications and commercial and economic aspects should be evaluated. Has your project an outcome that potentially could improve some community or group of people? Perhaps your project can impact on the lives of others for example In education, employment, health, public policy or services, security, the environment, general wellbeing etc. There may be commercial opportunities arising from your product or findings. Describe these and include how the project could eventually brought to deployment and to deliver value. Discuss the feasibility of doing that. It may be that your project could make some process more efficient. Try to quantify the savings or improvements, generally or in one or more scenarios. You should be realistic though and include the risks and any negative impacts of your work and the potential impact as well.

Your supervisor can guide you on what is appropriate, but typically the very best projects have shown results derived using scientific method, that could be publishable with little or no work or show an exemplary empirically based evaluation of a software product. Those projects will also fairly and honestly assess the potential impact of the work socially or economically.

Any publication of results of the student's work is left to the discretion of the supervisor, but you can expect appropriate credit to be given to your work.

1. Appendices

Appendices will not be marked but may be referred to by the assessor to aid their understanding. They are useful if there is something that helps in understanding earlier parts of the dissertation, but if included inline might break the flow or readability of the document. For example, there may be large tables of data, design documents, evidence of testing etc etc.