Volume 15 • Issue 4 • August 2021

Editor-in-Chief
Dr. Chen-Chi Shing

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

ISSN : 1985-1553

Publication Frequency: 6 Issues / Year

CSC PUBLISHERS
http://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

VOLUME 15, ISSUE 4, 2021

EDITED BY
DR. NABEEL TAHIR

ISSN (Online): 1985-1553
International Journal of Computer Science and Security is published both in traditional paper form and in Internet. This journal is published at the website https://www.cscjournals.org, maintained by Computer Science Journals (CSC Journals), Malaysia.

IJCSS Journal is a part of CSC Publishers
Computer Science Journals
https://www.cscjournals.org

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND SECURITY (IJCSS)

Book: Volume 15, Issue 4, August 2021
Publishing Date: 31-08-2021
ISSN (Online): 1985 -1553

IJCSS Journal is a part of CSC Publishers
https://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers, 2021

Dr. Sanaa Kaddoura
Department of Computing and Applied Technology, Zayed University
United Arab Emirates

Dr. Francesco Taglino
National Research Council
Italy

Dr. Rowanda Ahmed
Uskudar University
Turkey

TABLE OF CONTENTS

Volume 15, Issue 4, August 2021

Pages

EDITORIAL PREFACE

This is *Fourth* Issue of Volume *Fifteen* of the International Journal of Computer Science and Security (IJCSS). IJCSS is an International refereed journal for publication of current research in computer science and computer security technologies. IJCSS publishes research papers dealing primarily with the technological aspects of computer science in general and computer security in particular. Publications of IJCSS are beneficial for researchers, academics, scholars, advanced students, practitioners, and those seeking an update on current experience, state of the art research theories and future prospects in relation to computer science in general but specific to computer security studies. Some important topics cover by IJCSS are databases, electronic commerce, multimedia, bioinformatics, signal processing, image processing, access control, computer security, cryptography, communications and data security, etc.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Started with Volume 15, 2021, IJCSS appears with more focused issues. Besides normal publications, IJCSS intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

This journal publishes new dissertations and state of the art research to target its readership that not only includes researchers, industrialists and scientist but also advanced students and practitioners. The aim of IJCSS is to publish research which is not only technically proficient, but contains innovation or information for our international readers. In order to position IJCSS as one of the top International journal in computer science and security, a group of highly valuable and senior International scholars are serving its Editorial Board who ensures that each issue must publish qualitative research articles from International research communities relevant to Computer science and security fields.

IJCSS editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Scholar, J-Gate, Docstoc, Scribd, Slideshare, Bibsonomy and many more. Our International Editors are working on establishing good abstracting and indexing listing and a good impact factor for IJCSS. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can provide to our prospective authors is the mentoring nature of our review process. IJCSS provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Computer Science and Security (IJCSS)

Combining Approximate String Matching Algorithms and Term Frequency In The Detection of Plagiarism

Zina Balani *zina.0174810@gmail.com*
Department of Software Engineering
Koya University
Koy sinjaq, 44023, Iraq

Cihan Varol *cvarol@gmail.com*
Department of Computer Science
Sam Houston State University
Huntsville, TX 77341, USA

Abstract

One of the key factors behind plagiarism is the availability of a large amount of data and information on the internet that can be accessed rapidly. This increases the risk of academic fraud and intellectual property theft. As increasing anxiety over plagiarism grow, more observation was drawn towards automatic plagiarism detection. Hybrid algorithms are regarded as one of the most prospective ways to detect similarity of everyday language or source code written by a student. This study investigates the applicability and success of combining both the Levenshtein edit distance approximate string matching algorithm and the term frequency inverse document frequency (TF-IDF), thereby boosting the rate of similarity measured using cosine similarity. The proposed hybrid algorithm is also able to detect plagiarism occurred on natural language, source codes, exact, and disguised words. The developed algorithm can detect rearranged words, inter-textual similarity of insertion or deletion and grammatical changes. In this research three various dataset are used for testing: automated machine paragraphs, mistyped words and java source codes. Overall, the system proved to be detecting plagiarism better than the yet alone TF-IDF approach.

Keywords: Approximate, Hybrid, Plagiarism, Similarity, TFIDF.

1. INTRODUCTION

Onward entering the age of digital communication, the ease of sharing information across the internet has facilitated online literature searches. This carries the confidential risk of increased academic misconduct and educated attribute theft, as concerns about plagiarism increase [1]. Plagiarism is regarded a serious problem in modern research manuscripts [2]. This work will focus on two plagiarism areas: source code (also called "code clone") and natural language. Plagiarism in natural language is relatively hard to detect due to the rich morphological, syntactic, and semantic properties of natural language. Moreover, plagiarists use various paths to overcome plagiarism detection systems by replacing the original text using various types of hiding and using smart plagiarism techniques, including rewriting, synonym substitution, paraphrasing, text processing, text translation, and idea adoption [3]. To mitigate the negative effects of plagiarism, systems aim to identify cases of theft in paper or documents via comparing a large collection of documents with suspicious documents. Finally, systems detect if the suspicious document has been stolen or similar to each other [4].

There are two essential methods for automatic plagiarism detection: extrinsic / external plagiarism detection and intrinsic / internal plagiarism detection. Intrinsic plagiarism detection only analyzes the input document and finds some parts that have not been created by the same author without performing a comparison with an external corpus. External plagiarism detection requires a reference collection of documents that are considered as original. Suspicious documents are compared to all documents in this collection to detect duplicate or approximately matching components in the source document [5]. This study deals with external plagiarism detection methods.

Strings can be identical in two manners. There are cases where they share syntactically the same character sequence and cases where they have the same semantic meaning (such as synonyms). Similarity measures can break into four classes: character-based, q-gram, token-based, and mixed measures. Character-based and q-gram measures calculate similarity based on a sequence of characters in two strings. Token-based measures use white space, line breaks, or punctuation characters to break a string into words and symbols (called tokens), and calculate the similarity between two token sets. Mixed measures are a combination of character-based measures and token-based measures [6].

At the beginning of the 20th century, in the early 1970s, scientists have studied a large-scale copy of the program to prevent technology and software. There are several classic systems available for detecting plagiarism in program source code. Plagiarized program code is different than the plain language as different codes can perform the same function. Some smart plagiarists use certain methods to alter the code. For example, for loop becomes a while loop or adding large number of randomly generated intermediate variables [7].

In today’s developed technology where techniques of plagiarism detection are available, previous methods and techniques that were used to detect and find measures of plagiarism are no longer in use. Most investigators are now working on the success of hybrid algorithms to find the similarity.

Therefore, in this study the applicability and success of two specific hypothesis are investigated:

- Approximate string matching algorithms can be employed to detect plagiarism.
- Combining approximate string matching with TF-IDF will increase the plagiarism detection rate.

The proposed system can identify different types of plagiarism such as sentence reordering, source code, inter-textual similarity, and approximate copy similarity. Edit distance is used to find the similarity for mistyped textual information, if the obtained similarity measure exceeds the predefined threshold which is greater than or equal to 0.50. Once this threshold is reached, then TF-IDF counts the word as its original, which increases the frequency of the common words. As a result, the rate of plagiarism will be boosted which is measured by Cosine Similarity. The success rate of a hybrid version of approximate string matching and term frequency is investigated, and is the results reflected that a robust hybrid algorithm is capable to detect both plagiarism on natural language and source code (Code Clone).

2. LITERATURE REVIEW

A wide range of problems can be emerged when many students copy someone else's work and submit it as their own. To detect these cheats, the online learning management systems started to add plagiarism detection tools, and when two identical or sufficiently similar assignments are detected a flag is raised for the issue. However, plagiarists use a variety of methods to alter the submitted work to avoid detection by the system. In recent years, the problem of detecting plagiarism in natural language has attracted the attention of many researchers. A number of plagiarism detection tools have been developed specifically for English language.

According to [1], one of the oldest methods of detecting plagiarism was introduced by Bird in 1927, who investigated the application of statistical methods to detect plagiarism in multiple choice answers. Afterwards, methods developed in the 1960s pointed on detecting plagiarism in multiple choice tests. Levenshtein Distance (LD) was renamed after Vladimir Levenshtein, a Russian scientist who designed the algorithm in 1965, and is also known as edit distance. It's a measure of the similarity between two strings called source string (s) and target string (t). Distance is the number of deletions, additions, or replacements required to convert s to t [8]. This technique was used for near duplicate detection for several years. Back in the 1970s, researchers began working on similarity detection techniques for source code [9]. Since 1970s, numerous tools have been introduced to measure the similarity of code. They are used to address issues like code duplication detection, software license violations, and software plagiarism [10].

In 1975, Halstead suggested the first algorithm called the property counting method. The algorithm computed the operators and operands statistics seeming in the source program and used it as the main basis for determining the result [9]. Between the 1970s till mid-1990s was a golden period of developing different methods and algorithms of exact and approximate string matching algorithms [11]. Since the mid-1990s, the focus of research has shifted to the study of natural language text. In 1994 Manber suggested the concept of an approximate fingerprint. The basic rule is to measure the similarity among documents by string matching. This principle has been accepted by most researchers. Therefore, based on this, researchers increased word frequency counting, keyword extraction techniques, and accomplished matching by computing hash values for the text [9]. Experimental outcomes indicates that a hybrid technique that considers word frequency and character-based word similarity increases matching. The first venture in this direction was found in 2003 by Cohen et al. A measurement format named Soft-TFIDF, which extends the Jaro-Winkler method to combine the frequency weight of a word in a measure of cosine similarity and a measure of CLOSE at the character level [12]. In 2007, according to [1], Dreher proposed using a normalized word vector algorithm to calculate similarity based on VSM (Vector Space Model), which performs synonym generalization for each word. Despite the advent of more advanced approaches, paraphrase remained difficult. In 2012, Ekbal, Saha, and Choudhary suggested ways to detect plagiarism. This includes three main steps. First, the basic tasks of natural language processing are performed in preprocessing steps. The second step selects a set of source documents similar to the suspect document. A VSM is used to identify the source document for each suspect document. Finally, the similarity between the two documents is calculated using the cosine similarity. If the resultant measure of similarity exceeds a predefined threshold, the document is considered as a source document for the suspect document. In the third step, similar text is found in both source and suspect documents using the n-gram method [13]. In 2013, Investigators Combined both of VSM and Jaccard coefficient into one, the method fully utilizes the benefits of VSM and Jaccard coefficient, and it can extract more reasonable heuristic seeds in plagiarism detection. The preliminary outcomes indicate that the method can generate better performance. [14]. In 2018, a system was recommended for Urdu text, based on a distance measurement method, structural alignment algorithm, and vector space model. System performance is measured using machine learning classifiers (Support Vector Machine and Naive Bayes). Experimental outcomes demonstrate that the output of the suggested method is advanced compared to other existing model (i.e. cosine method, simple Jaccard measure) [15]. According to [16], the hybrid of the Rabin-Karp and Levenshtein algorithms provides a useful contribution. The degree of identity can be optimized so that documents can be classified precisely according to the content of the document. The hybrid algorithm is able to enhance the precision of the evaluators according to certain parameters, N-Gram, Base and Modulo. In 2021, AL-Jibory proposed an external plagiarism detection strategy based on a system integrating natural language processing (NLP) and machine learning (ML) techniques, as well as text mining and similarity analysis. The proposed technique uses a combination of Jaccard and Cosine similarity demonstrated by a design application used to identify plagiarism in scientific publications. [17]

The literature declared that till now the matching algorithms for detecting plagiarism are challenges. Therefore, the hybrid algorithms are advanced. Combining of Levenshtein Edit Distance LED and Term Frequency-Inverse Document Frequency TF-IDF behind of detecting plagiarized words, is also capable to enhance the rate of the plagiarism by detecting disguised words. Meantime, term frequency is implemented which is a powerful techniques can also detect the plagiarized source code from different programming languages.

3. METHODOLOGY

3.1 Edit Distance and TF-IDF Based Algorithms

• Levenshtein Edit Distance: One of the effective methods of comparing strings, two terms or mainly two sequences is the edit distance which calculates the cost of the optimal sequence of the editing process by adding, removing and replacing. There are multiple differences in the computation of the edit distance, while the Levenshtein distance proposed in 1966 being the most known in the literature. Distance is the minimum number of machining process that are converted from one string to another. Allowed editing operations are the adding, removing and substituting of individual characters [4]. Small discrepancies in the input data can still be pointing a plagiarism. This is why approximate string matching approach is used to cover this challenging area. A threshold between suspicious papers and the source document repository is established. If the predefined threshold is greater than or equal to 0.50, the word is considered similar to the original. Later, TF-IDF, which is explained below, is used to determine the frequency of the words to have a better understanding of the plagiarism rate.

• TF-IDF Weighting: TF-IDF is performed as an important determinant in text mining and information retrieval, which enables the establishment of a vector space where every vector indicates how a word is significant for a document in an aggregation through the combination of Term Frequency TF (t, d) and Inverse Document Frequency IDF (w) [18]:

a) Term Frequency TF it refers to the number of times a word is included in a document. Since the length of each document is various, a term may appear much more in long documents than in short documents. TF is defined as:

$$\mathrm{TF}_{(\mathrm{t,d})} = \frac{\mathrm{O}}{\mathrm{t}}$$

where O refers the amount of times that a term t occurs in a document, and t is the number of words in the paper.

b) Inverse Document Frequency IDF is a statistical weight performed to measure the significance of a word in a collection of text documents. It also has a built-in IDF functionality that decreases the weight of terms that appear regularly in the document set and increases the weight of words that appear infrequently. IDF is defined as:

$$\mathrm{IDF}_{(\mathrm{t,d})\mathrm{log}} = \frac{|\mathrm{D}|}{\mathrm{N}}$$

Where |D| is the complete number of documents and N is the number of documents with the term t.

c) Term Frequency-Inverse Document Frequency TF-IDF is computed for every word in the article by combining TF and IDF.

$$TF - IDF(t, d, f) = TF(t, d) * IDF(t, d)$$

TF-IDF algorithm is performed to detect the relevance to a query document and TF-IDF weighting for representing the percentage of terms in a document. Terms with high of TFIDF weighting indicates a powerful connection to the document, including the document query.

3.2 Vector Space Model with Cosine Similarity Measure

The vector space model (VSM) is one of the common techniques that uses the lexical and syntactic characteristics and describes the article in vector space. Several weighting schemes are then used for document presentations and comparisons. Term-frequency-inverse document frequency (TF-IDF) and term frequency-inverse sentence frequency (TF-ISF) are mainly used as two weighting schemes [19]. The TF-IDF weighting technique is frequently used in the vector space model along with similarity methods to find the similarity among two articles. Common measurements based on the vector space model are cosine similarity and jaccard coefficient, performed to represent text papers (usually every objects) as vectors in multidimensional space. In this study cosine similarity measure is applied.

The cosine similarity is a measure of the similarity that is computed by multiplying the cosine angles of the two vectors to be compared. The cosine 0 ° is 1, which is less than 1 to the value at any other angle. Thus, the similarity values of the two vectors are: If the cosine similarity value is 1, they are similar. This method is a traditional method that is often used and combined with the TF-IDF. The calculation of cosine similarity is implemented according to the below equation [20]:

$$Cos\alpha = \frac{A * B}{|A|*|B|} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sqrt[1]{\sum_{i=1}^{n} (A_i)^2} * \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

where A is the weight of every attribute of vectors A and B is the weight of every attribute in vector B.

Subsequently generating the vectors and converting the attributes in the vectors to weighted values, we can use cosine similarity measure to compute the likeness of those vectors. Basics of cosine similarity, the bigger angle shaped among two coordinate vector comparison papers, lower degree of similarity of papers. Moreover, the smaller degree of cosine similarity level means the similarity rate will be bigger [20].

3.3 System Architecture

The proposed system performs the following steps which are represented in figure 1.

1. Levenshtein Edit Distance Algorithm is implemented to find near identical information between the source and suspicious document. If determined threshold is greater or equal to 0.50 then the word is considered as plagiarized. Later, TF-IDF counts the word as it is original, which increases the rate of detecting plagiarism. Otherwise, the word considered as not plagiarized.
2. Cosine Similarity used as an efficient way to determine the similarity measures.

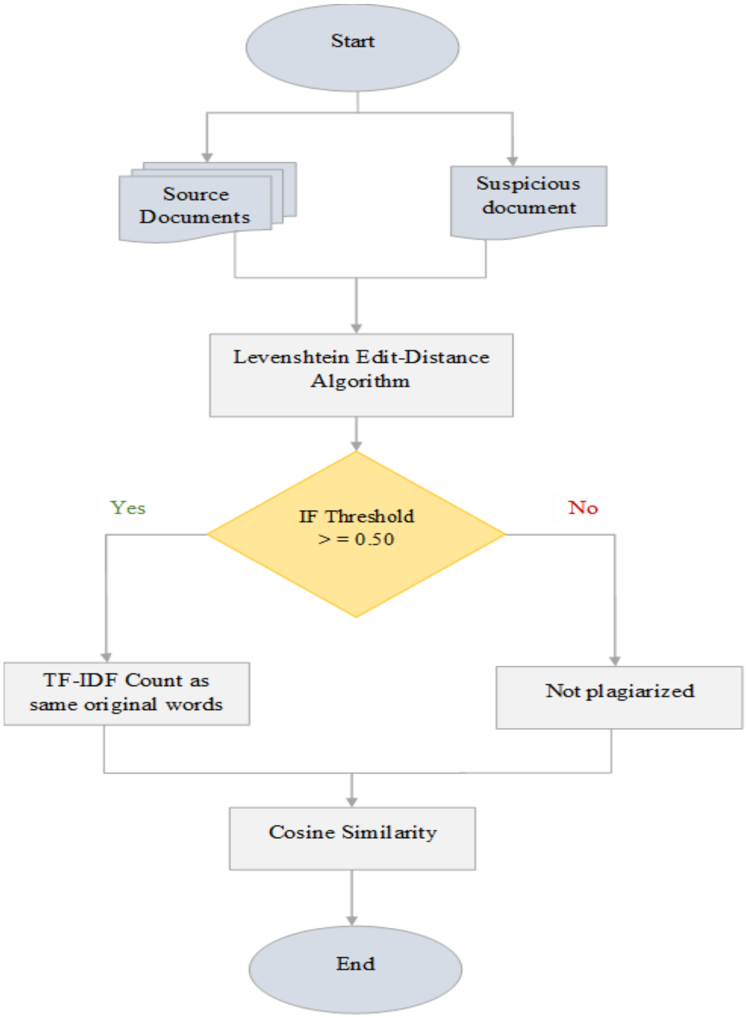


FIGURE 1: Proposed system flowchart for plagiarism detection system.

3.4 Implementation

Two techniques of matching algorithms have been successfully implemented to enhance plagiarism rate. Initially Levenshtein Edit Distance LED based on approximate string matching algorithms and Term Frequency Inverse Document Frequency TF-IDF based on exact string matching algorithms. Following, after finding Vector Space Model VSM, cosine similarity is implemented to determine the percentage of similarity between documents. Java programming language has been implemented in this project. And eclipse is used as a tool to execute the codes.

4. EXPERIMENTS AND RESULTS

The advanced plagiarism detection system is based on two similarity metrics. Accordingly, a percentage threshold is usually predefined in order to determine the cases of plagiarism to be found. Levenshtein edit distance has been implemented to determine near-identical similarity measure. If the percentage of plagiarism detected is greater or equal to 0.50 the system assumes that there is plagiarism and TF-IDF counts as it is a textual information identical to the one in original, which helps to identify near identical plagiarized information. Later, Cosine similarity measure is used to find the distance between the original and suspect document. Moreover, if the threshold is smaller than 0.50, the plagiarism will not be found. The outcomes demonstrates that the hybrid algorithm is more powerful than others. Because it can detect disguised or mistyped words in both of natural language and source codes and counts as its original which will boosts the rate of the plagiarism.

4.1 Datasets

This research contributes in providing a dataset for natural English language and source code plagiarism detection. The corpus contains several types of plagiarism cases including: simple copy/paste, word shuffling, and phrase shuffling, removed character, added diacritics, and paraphrasing. The corpus is expected to perform plagiarism detection approaches of available source documents comparable to each other. Plagiarism text files specific to this dataset are created for plagiarism purposes.

Three different datasets are used for testing: First, automated machine paragraphs [21] has the dataset size of 60.8MB which contains 79,970 documents, and it is divided into two sets: source documents (39,241) and suspicious document (40,729). Second dataset is Java source codes which were obtained from Githubs GH and stack over flows SO [22]. The dataset contains 180 java files, 97 of them are Github GH source code and 83 of them are on Stack Overflow SO. Finally, a Misspelled dataset [23] contains more than 6,000 misspelled words was used for testing. Table 1 demonstrate the result of our research after simulation test the optimum threshold is found to be 0.5.

No	Algorithms	Plagiarism Rate
1	Edit Distance	56.4%
2	TF-IDF	27.9
3	The Hybrid Algorithm	74.1%

TABLE 1: Comparison of plagiarism rate for algorithms .

4.2 Evaluation Measures

In order to evaluate the success of the result, three evaluation criteria are implemented in this study: precision and recall are two values that cooperatively are applied to assess the efficiency of performance of information retrieval systems. The F1 score is also a useful indicator for comparing two classifiers. F1 score produced by determining the harmonic mean of precision and recall. The results show that the version of the hybrid algorithm of approximate string-matching algorithms and term frequency is the most effective one which has a 0.851 F1 score, edit distance which has a score of 0.721 and TF-IDF is 0.436.

5. CONCLUSION AND FUTURE WORK

Edit distance provides an indication of similarity that may be too close in some cases. If user duplicates java code and performs a several alternates, such as, change of variable names, addition of two comments, the edit distance between the source and copy will be close. On the other hand, TF-IDF is the most commonly used weighting scheme for keywords to simplify all related articles. However, the current TF-IDF technique does not take into account the semantic correlations between terms, which can lead to a less relevant search for documents. Therefore, this study combined both the Levenshtein edit distance based approximate string matching algorithm and the term frequency inverse document frequency TF-IDF, thereby increases the score of similarity measured using cosine similarity. The robust hybrid algorithm is also able to detect plagiarism from both the natural language and source codes. Different forms of plagiarism can be detected such as (reorganization of words and inter-textual similarity of insertion / deletion). It can detect different types of plagiarism, such as added comments in Java source code or changes in the data fields and methods etc. In this investigation three different corpus are used for testing: automated machine paragraphs, mistyped words, and java source codes. From the results we find that the proposed algorithms are capable of detecting disguised and exact copy of articles which boosts the rate of the plagiarism detection.

In the future, we plan to extend the hybrid algorithm to be able to detect paraphrased text. We also plan to expand the scope of the source dataset with advanced programming topics (e.g. finding and sorting) and code files from other programming courses (e.g. object-oriented programming or algorithms and data structures).

6. REFERENCES

[1] M. Y. M. Chong, "A study on plagiarism detection and plagiarism direction identification using natural language processing techniques," 2013.

[2] S. Rani and J. Singh, "Enhancing Levenshtein’s edit distance algorithm for evaluating document similarity," in International Conference on Computing, Analytics and Networks, 2017: Springer, pp. 72-80.

[3] H. Cherroun and A. Alshehri, "Disguised plagiarism detection in Arabic text documents," in 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), 2018: IEEE, pp. 1-6.

[4] A. Zouhir, R. El Ayachi, and M. Biniz, "A comparative Plagiarism Detection System methods between sentences," in Journal of Physics: Conference Series. Vol. 1743. No. 1. IOP Publishing, 2021

[5] R. R. Naik, M. B. Landge, and C. N. Mahender, "A review on plagiarism detection tools," International Journal of Computer Applications, vol. 125, no. 11, 2015.

[6] N. Gali, R. Mariescu-Istodor, and P. Fränti, "Similarity measures for title matching," in 2016 23rd International Conference on Pattern Recognition (ICPR), 2016: IEEE, pp. 1548-1553.

[7] L. Qinqin and Z. Chunhai, "Research on algorithm of program code similarity detection," in 2017 International Conference on Computer Systems, Electronics and Control (ICCSEC), 2017: IEEE, pp. 1289-1292.

[8] X. Wang, S. Ju, and S. Wu, "Challenges in Chinese text similarity research," in 2008 International Symposiums on Information Processing, 2008: IEEE, pp. 297-302.

[9] X. Liu, C. Xu, and B. Ouyang, "Plagiarism detection algorithm for source code in computer science education," International Journal of Distance Education Technologies (IJDET), vol. 13, no. 4, pp. 29-39, 2015.

[10] C. Ragkhitwetsagul, J. Krinke, and D. Clark, "A comparison of code similarity analysers," Empirical Software Engineering, vol. 23, no. 4, pp. 2464-2519, 2018.

[11] K. Al-Khamaiseh and S. ALShagarin, "A survey of string matching algorithms," Int. J. Eng. Res. Appl, vol. 4, no. 7, pp. 144-156, 2014.

[12] T. El-Shishtawy, "A hybrid algorithm for matching arabic names," arXiv preprint arXiv:1309.5657, 2013.

[13] A. Ekbal, S. Saha, and G. Choudhary, "Plagiarism detection in text using vector space model," in 2012 12th International Conference on Hybrid Intelligent Systems (HIS), 2012: IEEE, pp. 366-371.

[14] S. Wang, H. Qi, L. Kong, and C. Nu, "Combination of VSM and Jaccard coefficient for external plagiarism detection," in 2013 international conference on machine learning and cybernetics, 2013, vol. 4: IEEE, pp. 1880-1885.

[15] W. Ali, Z. Rehman, A. U. Rehman, and M. Slaman, "Detection of plagiarism in Urdu text documents," in 2018 14th International Conference on Emerging Technologies (ICET), 2018: IEEE, pp. 1-6.

[16] A. H. Lubis, A. Ikhwan, and P. L. E. Kan, "Combination of levenshtein distance and rabin-karp to improve the accuracy of document equivalence level," International Journal of Engineering & Technology, vol. 7, no. 2.27, pp. 17-21, 2018.

[17] F. K. AL-Jibory, "Hybrid System for Plagiarism Detection on A Scientific Paper," Turkish Journal of Computer and Mathematics Education (TURCOMAT), vol. 12, no. 13, pp. 5707-5719, 2021.

[18] A. Mahmoud and M. Zrigui, "Semantic similarity analysis for paraphrase identification in Arabic texts," in Proceedings of the 31st Pacific Asia conference on language, information and computation, 2017, pp. 274-281.

[19] D. Gupta, "Study on Extrinsic Text Plagiarism Detection Techniques and Tools," Journal of Engineering Science & Technology Review, vol. 9, no. 5, 2016.

[20] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in 2016 4th International Conference on Cyber and IT Service Management, 2016: IEEE, pp. 1-6.

[21] Foltynek, T., Ruas, T., Scharpf, P., Meuschke, N., Schubotz, M., Grosky, W., Gipp, B. Detecting Machine-obfuscated Plagiarism [Data set], University of Michigan - Deep Blue, 2019.

[22] Baltes, Sebastian. Usage and Attribution of Stack Overflow Code Snippets in GitHub Projects — Supplementary Material (Version 2017-01-15) [Data set], 2018.

[23] Wahle, Jan Philip, Ruas, Terry, Foltynek, Tomas, Meuschke, Norman, & Gipp, Bela. Identifying Machine-Paraphrased Plagiarism (Version 1.0) [Data set], 2021.

Design and Implementation of a Predictive Model for Nigeria Local Football League

ADEBISI John *adebisi_tunji@yahoo.com*
Engineering/Computer
University of Namibia
Ongwediva, Namibia

ALABI Damilola *oluwadamilolaalabi1@gmail.com*
Engineering/Computer
University of Lagos
Lagos, Nigeria

Abstract

Sports prediction has become more interesting especially in the era of statistical information about the sport, players, teams and seasons are readily available. Sport analysts have opted out in their traditional ways of analyzing sport events and tends to leverage on the advantages of sports data; this enables more realistic analysis beyond sentiments. However, football game was considered in this research. Data from Nigerian Professional Football League (NPLF) was used to predict result based on different conditions such as home win, draw and away win of teams in the league. Machine Learning, k-Nearest Neighbor and mathematical Poisson distribution algorithm was hybridized using data mining tools together with Anaconda packages. The model accuracy was compared with other online bookmarkers, and it yielded 93.33% accuracy which will be helpful in making substantial profits in within the economy through the betting industries. This model is practically based on the home and away matches coupled with historical trends of goals scored and winning of previous matches, by implication, Nigerian football league will be more enhanced to catch up with their international counterparts and the players tends to get more feasibility from match result predictions for international participation and employment opportunities.

Keywords: Football, Sport, Machine-Learning, Poisson Distribution, Data Mining.

1. INTRODUCTION

Sport is defined as an athletic recreation which include tennis, soccer, and golf to mention a few. Any form of professionalism in the act is often called athlete. In simple term, sport can be either played indoor or outdoor. Sport can be further grouped into individual or team sport. During the sport competition, those who watch either in the stadium or on the screen are called fans[25]. While other watch for fun, some were paid to watch and are called the spectators. In this research, football sport is considered as it is the most populous type of sport in Nigeria. Football is a sport that is played by two different teams and each team with eleven players. Football game involved varying degree of kicking and nodding of ball with the main aim of securing a goal. The game has some rules to abide to [30]. Every player on the field will all participate to the overall performance of the match. The game is played on the field in a stadium with fans watching and cheering their favorite team. Officials also perform their duties one of which is the referee who watches the game closely and ensure that no rule is violated. Others may include commentators. The winner of the game is determined majorly by the team that secure more goals amidst other conditions based on the stage of the game in a tournament/league. Fans often try to predict the winner of football match which has led to the rapid increase of sports prediction [5].

Sports prediction has been increasing in popularity for many years in Nigeria especially in the era where everyone has access to internet connection. The urge to predict correctly among football fans gave rise to organizations specialized in sport prediction for profit making, hence the development of a more efficient and effective prediction model is required especially for local leagues. In the last few years, many stakeholders have opted to sport prediction and get rewarded for I t[32]. This has increased the participant in the sport industry as they are more passionate about the rewards, they get from predicting the outcome accurately [6]. Due to this, the sport industries provide the best reward and maintain relationship with their participant to make the sport more relevant in the society. In most society, when the stakeholder tries to forecast the event of a sport played, they go back and draft the past histories of the sport teams, then judged based on the information provided to predict whose team is going to win or lose. In case of sports like racing, tennis where the factors affecting the outcome of the game is less, it would be easy to analyze without sentiments, but when it comes to sport like soccer, several factors need to be considered before prediction. The prediction was usually carried out manually thus waste a lot of time since every data in the game that would determine the winning or losing team will have to be put into consideration. This led to the development of sport result prediction monitoring system using Nigerian Professional Football League (NPFL) as a case study.

NPFL is the highest level of the Nigerian Football League System (NFLS), for the Nigerian Club football Championships. It was organized by the League Management Company (LMC). Currently, the NPFL has twenty and four (24) football clubs all over the region in Nigeria who are currently playing in the league season. Although it was formally known as Nigerian Premier League. The team will play against one another, and the ultimate champion will be rewarded with the Nigeria FA Cup. The best three (3) teams qualifies for championship for Confederation of African Football (CAF).

With the upsurge in modern technology, many stakeholders use various tools as means to extract useful past data to evaluate them effectively. To overcome the problems faced by majority stakeholders, sport result prediction monitoring system is implemented. The model is incorporated into a web format for public access through the internet. Some merits of the monitoring system are; continuous support of training with new data for better accuracy, it also allows stakeholders spend less time analysis sport event outcomes. The manual methods of analysis sport results will void since the system will generate the result within second and hence decision can quickly be made by stakeholders to for result determination even when the game is ongoing. In addition is, awareness creation on the team involved which is better than just guessing the outcome of the game even without prior knowledge of such team. Increases in public interest on football game especially with prediction feature goes a long way in the improvement and monitoring of the local football league, then with the implementation of this model, there will be more awareness of the NPFL matches. Result of prediction will be uploaded online to enhance prediction accuracy.

1.2 Significance

In addressing the identified problem, this research designs and implement a predictive model for monitoring system using NPFL data. The objectives are to retrieve NPFL historical data from public database, design an interactive web platform that allow local sport fans and managers to predict and monitor NPFL sport match outcome and implement the designed model. The implication of this research is to monitor sport event outcome using the NPFL. This will enhance sport managers victory strategies and further strengthen sport fans to watching more of this local league since this system predicts the results of each game for sport fans and in return get rewarded for. This brings a lot of advantages to the local league management with easy accessing of this predicting platform, more accuracy implies more awareness to the public and in turns give popularity to the leagues. This model will also maximize data retrieved during matches as it gets popular through regular update thus, boosting the online community of local sport fans and this can gain international recognitions.

1.3 NPFL Peculiarities

The draw of NPFL consist of two groups; Group A and Group B. each group consists of twelve (12) teams. The teams are paired up in their respective groups and will play against one another in their groups both in the home and away; by the end of the tournament, a team would have played against other twenty-three (23) teams in home and away each to complete a total match of (46). Points will be awarded according to the win draw or lose. For a win match the team will be awarded 3points, for a draw match team will be rewarded 1point, and lose with 0point. The best three (3) teams in each group qualifies for the next stage and the four teams with the lowest points in each group relegates (out of the tournament) and demoted to Division two (2) considered in determining the winner or loser. Rules in global football tournament also hold in NPFL.

2. THEORECTICAL FRAMEWORK

There are number of events that can be predicted in a football match such as the number of goal scored either for half time or full time, the player to score the next goal, the team to score first, or even to predict a fix match which is a determination of the exact goal that will occur in the match [9]. Nevertheless, football is full of unpredictable event and so this work will only be classifying our prediction into the win/drawn and lose. The research interest is in the supervised learning method of Machine Learning (ML) and in particular the classification methods such as Logistic Regression (LR) Model, Decision Tree (DT), Random Forest (RF), k-Nearest Neighbour (k-NN), and Support Vector Machine (SVM)[37]. Others include Artificial Neural Network (ANN), Lazy learning and Bayesian method. Machine Learning capacity is explored in this research alongside with others for effective prediction. ML allows computer decides by itself with little or no assistance. ML uses data and statistics for analyses just like statistics but uses different methodology. The major two categories of ML is the supervised and unsupervised learning. The Supervised learning is later grouped into regression and classification method. Some of most important ML methods are: “Supervised Learning (Classification and Regression) [22].

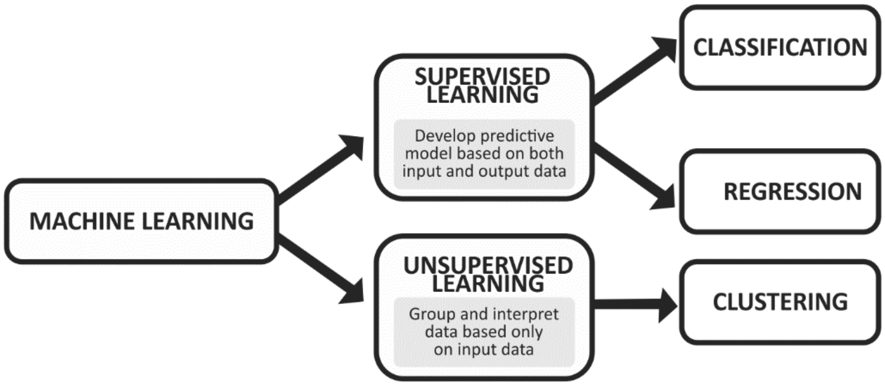

MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

FIGURE 1: Supervised and Unsupervised L1earning (Source: https://de.mathworks.com/).

a) Artificial Neural Network

ANN mimic the human brain; the network has neurons. Each neurons have a weight value and are connected at the node. The network is consist of at least one input and output with some interconnected neurons in between the input and output nodes [4]. The core of neural networks, neurons, are just simple activation function that has multiple inputs and one output. The neuron can be seen as a composition of several other weighted neurons and the network can be described by the network function in equation 1.0 as described by [22]. The neuron output may be the input of another neuron and every neuron weight the input and calculate its activation value give;

$$f(x) = k((\sum_i M_i c_i(x)) \qquad (1.0)$$

Where *Mi* are weights, *Ci* are other functions and K is the activation function. Generally, ANN continuously changes the weight of each hidden node depending on the output weight.

b) Bayesian Method

The Bayesian model is one of the well-known supervised machine learning classification techniques. It is easy and effective in performing well with unrelated and distinct features. Bayesian classifier is a probabilistic forecast system that implies that all characteristics variable involves does not in any way depend on one another i.e. are independent from the class variable. There are some separate characteristics in each category. Based on prior information, it then predicts the outcome event information. In the existence of complexity and uncertainty, Bayesian networks are graphical models for estimation [6]. Bayesian network's primary concept is effectively derived from Thomas Bayes' called the rule of the Bayes. The Bayes' theorem is represented in Equation 2.0 or Equation 3.0 by[28].

$$\text{P(hypothesis | evidence)} = \frac{\text{P(evidence|hypothesis).P(hypothesis)}}{\text{P(evidence)}} \qquad (2.0)$$

Where:
P(hypothesis | evidence) The likelihood of the hypothesis after proof is observed.
P(evidence |hypothesis) Describes the likelihood of evidence for a given hypothesis
P(evidence) as the evidence of the unknown cause in the event.
P(hypothesis) The probability of all event before observing their effects.

Generally, the formula can be interpreted in the equation below.

$$P(J|K) = \frac{P(K|J).P(J)}{P(K)} \qquad (3.0)$$

Where:
P(J) is previous likelihood of J.
P(J|K) is the later likelihood of J given K
P(K|J) is Conditional likelihood of K given J
P(K) is the prior probability

c) Linear Models and Logistic Regression

Linear models are collection of regression techniques that assume that the output figure is a linear mixture of all input variables.

Consider the diagram above which is the graph of a straight line

$$y = \alpha + \beta x \qquad (4.0)$$

In the straight-line Equation in 4.0, β is the slope and α is intercept on y. This relationship may not be true for large dependent and independent variable which lead to another equation when observing n sample of data as shown in Equation 5.0.

$$Yi = \alpha + \beta x_i + \varepsilon_i \qquad (5.0)$$

The aim of this model is to determine the value for the $\alpha\ and\ \beta$ for which the output is formed on the best fit line on the other way, Logistic Regression (LR) is a distinguished classification method. Unlike linear regression, LR relies on linear feature mixture, which is then plotted by the logistic feature to a value between 1 and 0. Thus, dependent factors should have a constant significance that, in turn, is a function of event probability. There are two phases of logistical regression. First, estimate the probability of each group's characteristics and second, determine the cut-off points and categorize the characteristics appropriately by [23].

d) Decision Tree and Random Forest

DT is a common ML method for linking entry factors (input) depicted in the branches and nodes of the tree with an outcome value (output) displayed in the leaves of the tree. Trees can be used either in classification analysis, by producing a class tag or in regression analysis, by producing an actual number. Decision Trees can be installed using various methods, including the most common CART or ID3 DT systems. However, DT can often become incorrect, particularly when subjected to big amounts of information from practice as the tree becomes a victim of over fitting. This happens when the model fits the training information but cannot generalize to unforeseen data. Random Forest (RF) on the other hand is a combination of different DT in DT training output node is the input to another DT which form the RF classification. RF has better performance when dealing with over fitting.

e) Support Vector Machine (SVM)

SVM are both classification and regression ML algorithm. An SVM system reflects the training data as space points. New variables are plotted and categorized as the class they drop into (which becomes part of the hyper plane) in the same manner as the training data. The kernel trick can be used if the information is not linearly separable by using various feasible kernel features such as (RBF) or nonlinear features. The SVM algorithm looks for an ideal hyper plane that functions as a border of choice between the two categories. While training, SVM lasts longer than other techniques. The algorithm is highly accurate due to its elevated capacity to build non-linear, complicated choice boundaries.

2.1 Lazy Learning

It is a ML technique which has no actual model for the training. The overall model is trained based on the new input. It is best used for data set that has relatively small features but has a reasonable large amount of data. One of the algorithms that can be classified as Lazy learning is the k-nearest neighbor which can be used for regression and classification analysis. The k-value depends on the sets of data set. The larger the value of k the lesser the noise effect. In this model if the new data set (green object) is to be classified.

a) Poisson Distribution and Poisson Regression

This is used in the statistics field to determine the outcome of an event in form in probabilities. The Poisson model was first described by Simeon Denis Poisson who was a mathematician at that time in France named Haight in 1967. [11]. This model is used to determine the occurrence of an event that takes place in interval. Example where this model can be applied is in football game to check the goal probability that can occur in a match between two team using the past average goals scored by individual team. Using the mathematical model f (k; λ) [10].

$$\text{PMF} = \Pr(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}\,;\ \lambda > 0 \qquad (6.0)$$

Equation 6.0 is a Probability Mass Function (PMF). Where λ is the avg. number of event that can occur i.e. the value of X in the conditional variable event $\lambda = E\ (k|X) > = 0$ and the e is the Euler's number of 2.71828… and the k ϵ {0,1,2,…} is a positive integer in factorial. The Poisson Regression is used to generalize the Poisson distribution into a linear model.

2.2 Empirical Review

Various researchers have come with different methods and techniques to predict result. Meanwhile, the analysis serve as a guide line for the public interested in football both for the coaches viewers. In this section, this research consider various researchers who have done relevant works in this area. Graham and Stott[9] applied an ordered probit-model using one fixture in the team to determine the strength of the team individually, but the major drawback in this model is lack of dynamic update.

[2] developed a machine learning framework to expand areas of necessity for good predictive accuracy in sport prediction based on artificial neural network to formulate informative strategies

however not designed for football. [30, 31 and 32] presented a comprehensive overview of big data management which is not really in sport but some of its concepts are important to this work in social data area since sports are regarded as social activities. Impact of software architecture in product upgrade and maintenance become very important as system deployment spanning across decades risk increased complexity that could only be managed by proper maintenance, which is very useful to enhance the longevity of the work examined by this research. Exploring the power of real time result processing and application is the works of [1], an aspect of their methodology during data gathering was very useful for this research to ensure real time prediction result processing. [37] uses MATLAB for sport prediction unlike the web application developed for this work. The works of [35] and [36] applied machine learning to the prediction of the outcome of professional sports events and to exploit "inefficiencies" in the corresponding betting markets. Tenis was the major subject of discuss and not football as addressed in this research.

[27] predicted results in the National Football League (NFL) using an ANN model which was conducted during one of his initial studies. He selected five fixtures in the first eight rounds of the league, consisting of “yards gained, rushing yards gained, turnover margin, time of possession and betting line odds”. The research used unclassified part of ML to determine which team is best and which one is poor. Achieved 61% accuracy. However, limitation of this study was the limited number of fixtures, and the model cannot be used to classified match for a win/lose/draw.

[29] developed a model using Artificial intelligence hybridized with Multiple Linear Regression and expert human predictions to determine the outcome of soccer and rugby game. The English Premiership Football teams and 2 Premiership Rugby Union team were used as a case study however much details of the uniqueness and relationship between soccer and rugby was not presented unlike, McCabe and Trevathan[19] who presented an extension of earlier work after three years Reed and O’Donoghue[28] published theirs. Artificial intelligence was used for prediction of sport game event. A multi-layer perception was used to model the system. “The information used in this model has been drawn from various sources and includes four main sports in the league which were the Australian National Rugby League (NRL), the Australian Football League (AFL)” McCabe and Trevathan[19] explored Super Ruby and English Premier League Football (EPL) in 2002. The fixtures acquired were focused exclusively on information such as score line, latest results and "league ladder" location compared to other teams. An accuracy of 65.1%, 63.2%, 54.6% and 67.5% AFL, NRL, EPL, was obtained and Super Rugby League, respectively. The work tried to decrease the Bayesian hierarchical model's over-shrinking difficulty by implementing a blend model, making the system more complicated and time consuming. The system can forecast outcomes for only a team.

[33] implemented a system for predicting sport game. Data set of two successive seasons of the (NBA) League were used. Data were collected from NBA league then uses module in his system. Unfortunately, the system uses the referent classifier and thus, absence of comparison with similar research in which to compare the predictions on the same set. Also the system uses manual fixture selection, however, the prediction accuracy is quit reliable. Furthermore, a group of researchers also worked in the area of the National Basketball Association (NBA) sport in 2010, Miljkovic *et al* uses data mining to forecast the results of NBA league basketball matches. The model uses classification problem which include the native Bayes method. Also multivariate linear regression to determine NBA spread. The data set of 2009/2010 NBA season was used to evaluate their system and achieved an accuracy of 67%. The system only uses win/lose as required in NBA, so it cannot be used for sport competitions that give room for draws such as football.

[18] considered the Problem in the competitive horse racing framework and show how to adapt the RF Classifier to forecast the results. The assessment was focused on a dataset of 1000 games between 2005 and 2006 at Hong Kong racetracks. The major drawback in this method was the complexity of the model used for a simple sport as it win/lose is mostly determined by individual/ horse capability. [15] presented the application of the weighted probability strategy

using weighting systems that are simple to obtain. This method focuses on the amount of goal the two competitors secured. The data from the Champions League was used to show the capacities of the suggested approach. Although, this strategy makes it possible to predict the initial score adequately and accurately, it does not account for big or unexpected final score that may deteriorate parameter projections. It uses the goal scored by team; it does not give account for shocking goals that can weaken its approximations.

[4] suggested a predictive scheme for football games beating the likelihood of bookmakers (odd). The prediction for the matches' uses previous result of the team involved as a guideline. The match projections use the prior team outcome as a guideline. Data set of the last 15 years for the Dutch football competition were used. In their inquiry, some of the most significant ML algorithms were used such as the BN and a form of LR. Features such as number of home wins, number of goals etc. The accuracy was not reliable since it was never much higher than a mere 55%, also, there is absence of comparison with similar research. In the works of [24], the approach to forecast results of soccer matches using the NETICA software. The Spanish League-Barcelona team during the 2008-2009 season was used to test the performance of the technique. Factors which affect the outcome of football matches were evaluated which were divided into non-psychological factors such as average player's age, history of five previous matches and psychological factors like the weather. The number of goals conceded by each team was categorized by BN which determines the (win/lost/draw). Following the results; a comparison with 2008-2009 seasons and gives an accuracy of 92%. The limitation of this model was football data affected the outcome of match in terms weighting. Although, many factors were considered in the model but most of the factors has little effect on football outcome as used.

[11] focused on data mining techniques used for sport prediction. The research work reviewed various techniques such as ANN, Decision trees, SVM, Fuzzy Systems, Bayesian methods among others. They evaluated available literatures in this regard and detected two major challenges. In this respect, they assessed literatures and identified two significant difficulties. First, the small precision of projections demonstrated the need for further studies to achieve accurate result. Second, the absence of an extensive collection of statistics pushes the research to gather information data from sports pages". They propose a range of alternatives to address these problems one of which was to improve prediction accuracy through ML and data mining techniques that have not been used in the field of football prediction but have been used in other field and yielded good results. They concluded that the application of hybrid algorithms can increase prediction accuracy.

[12] uses data mining tools to implement the model and weigh and predict the outcome of a soccer game. Using data mining software such as Rapid Miner to mine football information. Fixtures like the number of goals scored, moving average of teams, performance within a season, players and managers' performance indices, history of team, and weather conditions were used in this research. The system uses nine fixtures with optimization of weight. The structure utilizes two distinct methods of information mining which were ANN and LR techniques. The data set comprises of 110 games in the 2014-2015 season of the English Premier League. Comparing the result outcome of their model, "a greater forecast precision was obvious when weighting optimized characteristics. The ANN technique yields 85% while the LR yield 93 %. Although, the LR model cannot predict if a match must be draw. In this case the ANN is more accurate when compared with LR technique when predicting if a match must end at draw. The major limitation with the ANN approach is that it requires major factors that affect the outcome of a match and need as many data as possible to be more accurate.

[2] did something quite similar to Dixon and Coles[5]. They proposed an advanced modeling strategy and predicted the result-based Poisson Auto-regression with exogenous covariates (PARX) of football games. Used the 2013/2014, 2014/2015 and 2015/2016 English Football Premier League information season. This research work too advantage of the goal intensity feature to determine the best team. Based on the performance of the model it yielded of 43.27%, 44.96% and 12.63% respectively for the seasons 2013/2014, 2014/2015 and 2015/2016. The

threshold was modified to 0.3 and they achieve a return greater than 87% for those three Premier League Season. The limitations behind this model is that sophisticated features and factors that determine outcome of football was not considered. Considering the fact that the accuracy was low and thus depends on threshold value which may not be determined in real life event. A return greater than 87% was achieved for those three Premier League Season with limited fixtures and factors that determine outcome of football.

[32] used hybridized ANN and Linear Regression to predict the score outcome for matches played in the Spanish La Liga over five seasons. Features were gotten from FIFA 18 game database. They were able to achieve 71.63% accuracy with LR. Their ANN model achieved an accuracy of 63.1% from the match history database and 69.2 percent from the combination of the match history in the Team database. The major limitation was that the ANN model takes time and also large and sophisticated data is needed for this approach. [31] presents the use of the Google Prediction API to analyze prior cricket game information and predict cricket match outcomes. System of predictions works on the principle of machine learning which uses Regression Algorithms and Classifiers. The India team and other teams were used as a case study. Supervised machine learning was used. The proposed model yielded outcomes depending on earlier data supplied. The more the system of data is trained, the more superior outcomes. The test information was used to verify the forecast precision between India and other teams with a total 9 out of 10 games properly predicted. However, the limitation to this model was that it requires a lot of data for better accuracy, if the data is less, the accuracy is also less.

[8] predicted soccer match outcome based on the chances of bookmakers by using k-nearest method using the super league of Turkey competition 2015/16 season. The neighboring k-nearest algorithm was chosen as the assessment method, used the estimated results were compared with the bookmaker as a reference. Their model depends on the bookmarkers' odd. It was observed that there may be inconsistency in result if the bookmarker prediction is inaccurate. [29] applied fuzzy predictive classifier and proposed a model to predict the Brazilian football match in local league and. It forecasted 71 of 97 (73.2 percent precision) victories and 21 of 44 (47.73 percent precision) losses. The Maximum Likelihood classifier estimated, however, only 9 comes out of 49, a poor precision of 18.37 %. It is limited to first order uncertainty.

Having reviewed quite a number of related works to this research, the uniqueness of this work lies in the geographical area of application with the peculiarities of the game administrative pattern obtainable in Nigeria. One of the major gaps bridged by this work is the gap between the international league and local football leagues. In practical this work offers a more reliable approach that is best suitable to local league football result predication. In addition is the combination of methods as discussed in the subsequent section of this work.

3. METHODOLOGY

The method used in this research are encompassing, most of which are deductive in nature. This method aims at leveraging existing approaches and apply some of its specifics towards achieving the set objectives of this work. This work used an hybrid of k-NN model for data related to goals and Poisson distribution/regression mathematical model for other information. It centers on how effective prediction of the Nigeria football league can be achieved using fixtures from past matches. The importance of match fixtures when it comes to prediction becomes inevitable, event outcome of the matches played, among others. Algorithms were formulated to calculate the probabilities of match outcome. The methodology comprises of four (4) modules.

a) Data Collection: This enhances the model for effective and accurate prediction, outcome of sport results. Data were collected manually and electronically from valid local sport website peculiar to the Nigeria League in line with the objectives. The data in *.csv/.xlsx* format for easy manipulation during implementation. NPFL sport data were collected from the official website and other secondary sources as shown in Table 1.0. A modified database was created from the public version for the purpose of this research and consolidates information from four distinct sources that to suite its method. Below is sample raw data from NPFL website for 2017 league showing some highlight of all clubs in season 2017.

b) Fixture Selection: This module determines the most important fixtures. Data mining tools were used i.e. a combination of machine intelligence with human perception. The data available are majorly based on the goals scored between two teams in the league as obtained from www.rsssf.com/tablesn/nigchamp.html [last accessed 1/10/2019]. As a result of this, the goal scored were analyzed and used in this prediction model. Several indicators for the home team and for the away team were created. Output result FTR (Full Time Result) will be (H-Home win, D-Draw, A-Away win). Hence, data were separated into training and testing.

c) Classifier/Algorithm Selection: The algorithm used in this research solely depend on the data fixture which is majorly the goals scored in the match between two teams. Therefore, the best classifier for this dataset uses k-NN model. Although, the goal scored were analyzed with the Poisson distribution and regression mathematical model. Predictions were classified into (Home Win- "1", Draw-"X" and Away Win "2") which was the scope of this research. For effective result, these classifiers were hybridized since hybrid model has proven effective in past research.

d) K-NN Algorithm Analysis

The k-nearest neighbor was implemented due to the nature of data available. The k-nearest method uses Euclidean distance, (x_1, x_2) which find the closest distance between two data sets and classified it based on the value of k.X_1 is the target distance value while X_2 is the data to compare. Generally, the Euclidean distance is given by;

$$||X_1 - X_2|| \tag{7.1}$$

For probability of the Home *Win*, *Draw,* and Away *Win* is given as

$$\varphi = \varphi_{W,}\ \varphi_D,\ \varphi_L \tag{7.2}$$

Therefore, the probability of the class is between 0 to 1. The percentage of the probability can be

obtained as $P_{Per} = \frac{\varphi_i^{-1}}{\sum_{i=1}^{n} \varphi_i^{-1}}$ (7.3)

Generally, using the Euclidean distance formula, a standard deviation can be obtained for each classifier using 7.4 Murphy.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_1 - \overline{x_2})^2} \tag{7.4}$$

Where 'n' represents the total number of team features to be determined i.e. PTS, W, D, L, GF, GA. Detailed analysis shown in table 1.0. Using dataset from 2009 to 2019 of NPFL, the Points were calculated for Points (PTS), Won (W), Drawn (D), Loss (L), Goals for (GA), Goal against (GA) and Goal Difference (GD). For each variable, the average performance for the points (PTS) were calculated. The average point for the team from 2009 to 2019 (10 years) is given by $\frac{1}{n}\sum_{i=1}^{n} x$ where x is the variable to determine and calculated. Same for other parameters. After this has been determined for each team the Euclidean is calculated.

The formula is given as $\sqrt{\sum(x_1 - x_2)^2}$ (7.5)

Where x_1 and x_2 is the Euclidean distance where x1 is the parameter of the first team and x2 is the parameter for the second team to be compared.

e) Analysis of Goal Scored with Poisson Distribution

With Poisson distribution and regression formula, odd in percentage of home win, draw and away win for each team are determined respectively. Previously identified equations are applied to predict the outcome of football matches using the goals scored during the match. The variable in the equations are be determined from the scope of the data. Football team is measured on the basis of the "**Attack/Defence Strength**" which is in this instance determined using the goals scored in general. Where the attack strength is the ability of a team to score and the defence strength is the ability for a team concede goal.

$$\text{Attack Strength of NPFL} = \frac{\text{total goal scored in the NPFL}}{total\ number\ of\ match\ played}. \qquad (8.0)$$

$$\text{Defence Strength of NPFL} = \frac{\text{total goal concedes in the NPFL}}{total\ number\ of\ match\ played} \qquad (8.1)$$

In this case, it involves two teams which is either **Home** or **Away** donated as (H or A). Equations 8.0 and 8.1 are used to predict the probability of the Home Win, Draw and Away Win of a team basically using the illustration of two teams **Team "A"** and **Team "B". Team A** is **Home** while **Team B** is **Away**

Attack Strength of **Team "A" -Home =**

$$\frac{\text{total goal scored in the Home}}{total\ number\ of\ Home\ Match} \div \text{Attack Strength of NPFL} \qquad (8.2)$$

Defence Strength of **Team "B"-Away =**

$$\frac{\text{total goal away conceded from Home}}{total\ number\ of\ Away\ Match} \div \text{Attack Strength of NPFL} \qquad (8.3)$$

Equations 8.0 – 8.3 are used to determine the possible goals that **Team "A" - Home** is likely to score in NPFL which is given as;

Attack Strength of **Team "A** × Defence Strength of **Team "B" ×** Attack Strength of NPFL - (8.4)

The same steps were applied to determine the goal **Team "B" –Away** is likely to score.

Attack Strength of **Team "B" -Away =**

$$\frac{\text{total goal scored Away Match}}{total\ number\ of\ Away\ Match} \div \text{Defence Strength of NPFL} \qquad (8.5)$$

Defence Strength of **Team "A"-Home =**

$$\frac{\text{total goal Home conceded}}{total\ number\ of\ Away\ Match} \div \text{Defence Strength of NPFL} \qquad (8.6)$$

The above equations can be used to determine the possible goal that **Team "B" - Away** is likely to score in NPFL which is given as

Attack Strength of **Team "B** × Defence Strength of **Team "A" ×** Defence Strength of NPFL (8.7).

The Attack/Defence Strength value is the "λ"for the Poisson equation as in equation 2.5 Pr(X = k) $= e^{-\lambda}\frac{\lambda^k}{k!}$ $\lambda > 0$

ISSN: 1985-1553, https://www.cscjournals.org/journals/IJCSS/description.php

The value of k is the goal and it is a positive integer where k = 0,1,2,3… The distribution gives the probability of number of goals scored by individual team. Since both goal are not a dependent variable i.e. they do not depends on one another in the match.

Table 2: Probability Distribution of Goal Scored (K = 0 To 5)

Goals	**0**	**1`**	**2**	**3**	**4**	**5**
Team A	$P_A\{0\}$	$P_A\{1\}$	$P_A\{2\}$	$P_A\{3\}$	$P_A\{4\}$	$P_A\{5\}$
Team B	$P_B\{0\}$	$P_B\{1\}$	$P_B\{2\}$	$P_B\{3\}$	$P_B\{4\}$	$P_B\{5\}$

Source: https://www.pinnacle.com/en/betting-articles/Soccer/how-to-calculate-poisson-distribution/MD62MLXUMKMXZ6A8.

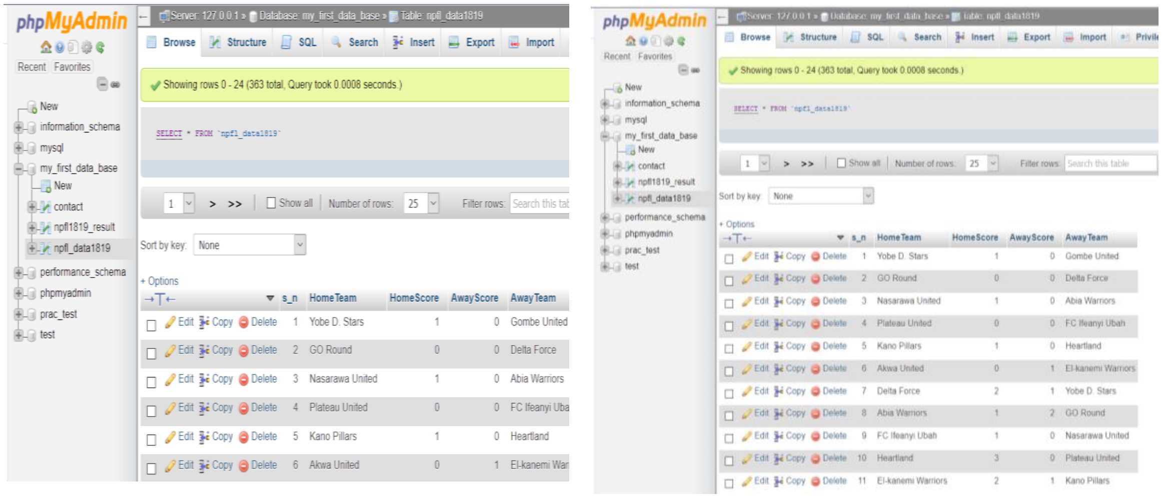
By implication, if we are to determination of a draw outcome between Team A and B, will be deduced from the summation of Joint probability of equal goal i.e. $\sum(A \cap B)$ which is $P_A\{0\}$. $P_B\{0\}$ + $P_A\{1\}$. $P_B\{1\}$ + $P_A\{2\}$. $P_B\{2\}$……..+ $P_A\{5\}$ $P_B\{5\}$. This can be computed using the Poisson regression for various outcome of event that we are interested in.

$$\Pr(X = k) = \frac{\exp(-\exp(k'\beta))\exp k\prime\beta^{x}}{x!}. \qquad (8.8)$$

4. IMPLEMENTATION

a) Tools and Database System Used

Popular data science tools were considered for this research i.e. Anaconda with python libraries. Python programming language Data was migrated into the database from the raw environment after which the model was used to build the classification. Various inbuilt libraries such as the sklearn, panda was used to build the model. The result of the model was stored in a csv/xlsx file and later displayed on the web through the database (MySQL) used for web analysis of result. A database driven interactive webpage was implemented as shown in Figures 2 and 3, and data stored in the database are displayed on the platform using the Django/Flash framework. These frameworks performed optimally with the python language and delivered a dynamic website. The interactive and dynamic nature of this web application make this work better and unique compared to the related work in this research.



Showing rows 0 - 24 (363 total, Query took 0.0008 seconds.)

```
SELECT * FROM `npfl_data1819`
```

s_n	HomeTeam	HomeScore	AwayScore	AwayTeam
1	Yobe D. Stars	1	0	Gombe United
2	GO Round	0	0	Delta Force
3	Nasarawa United	1	0	Abia Warriors
4	Plateau United	0	0	FC Ifeanyi Uba
5	Kano Pillars	1	0	Heartland
6	Akwa United	0	1	El-kanemi War

(a)

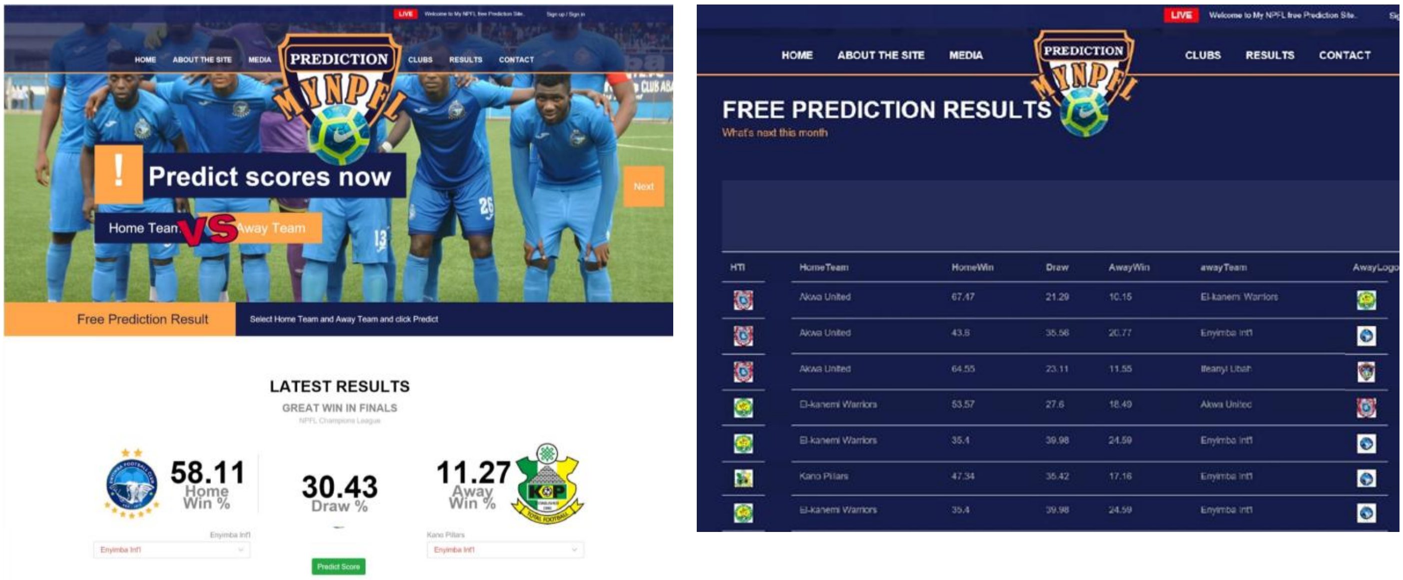
Showing rows 0 - 24 (363 total, Query took 0.0008 seconds.)

```
SELECT * FROM `npfl_data1819`
```

s_n	HomeTeam	HomeScore	AwayScore	AwayTeam
1	Yobe D. Stars	1	0	Gombe United
2	GO Round	0	0	Delta Force
3	Nasarawa United	1	0	Abia Warriors
4	Plateau United	0	0	FC Ifeanyi Ubah
5	Kano Pillars	1	0	Heartland
6	Akwa United	0	1	El-kanemi Warriors
7	Delta Force	2	1	Yobe D. Stars
8	Abia Warriors	1	2	GO Round
9	FC Ifeanyi Ubah	1	0	Nasarawa United
10	Heartland	3	0	Plateau United
11	El-kanemi Warriors	2	1	Kano Pillars

(b)

FIGURE 2: MySQL table of showing the database structure for experimentation dataset.



HTl	HomeTeam	HomeWin	Draw	AwayWin	awayTeam	AwayLogo
	Akwa United	67.47	21.29	10.15	El-kanemi Warriors	
	Akwa United	43.6	35.56	20.77	Enyimba Int'l	
	Akwa United	64.55	23.11	11.55	Ifeanyi Ubah	
	El-kanemi Warriors	53.57	27.6	18.49	Akwa United	
	El-kanemi Warriors	35.4	39.98	24.59	Enyimba Int'l	
	Kano Pillars	47.34	35.42	17.16	Enyimba Int'l	
	El-kanemi Warriors	35.4	39.98	24.59	Enyimba Int'l	

(a) (b)

FIGURE 3: Web application landing pager and the result prediction Interface.

4.1 Result Analysis of NPFL

Data of NPFL 2018/2019 contains 380 matches in the league season. The data frame in the columns are the team names and the number of scored goals for the home and away team as shown respectively. From the bar chart analysis in figure 10, the home teams occupies majority of the graph which means that home teams have more advantage of winning probability. Usually in football, away teams travel and may be more fatigue than their opponent in home teams. Also, the away team may not be familiar with the football pitch. This is generated from the implementation code. As a result, the highest frequency is home team winning, they tend to score more goals on average which means that home team scores more goals in the league which was because of the home advantage as stated earlier. It is emphasized the use of Poisson distribution since the goals scored describe the outcome of the match and thus the number of goals scored during the match is independent of the duration the match has commenced. This also gives the clue that home team secure more goals than the away team. Goals scored are better expressed independently by finding the goal distributions. Equation 6.0 and the Poisson model was used to generate the distribution. Figure 4; this shows predictions of the number of goals per match in NPFL 2018/2019 season. Furthermore, the actual data sample compared with the Poisson distribution model is useful for comparison match prediction.

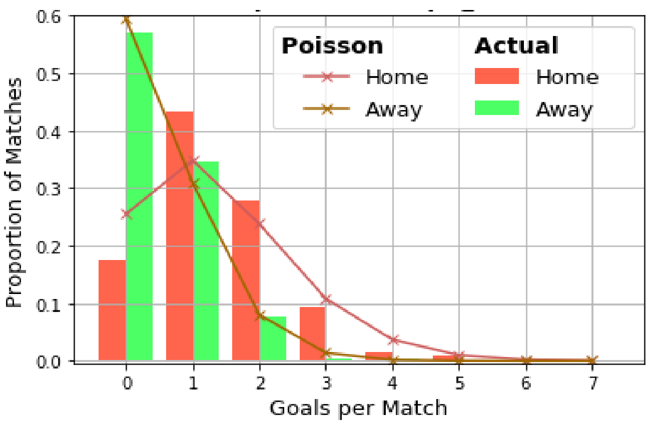


FIGURE 4: Poisson distribution of goals per match (NPFL 2018/2019 season).

4.2 Model Output Result

The test data consist of 24 teams each team plays 23 games for both the home and the away. The accuracy of the model mostly depend on the previous games. In this case, all the teams were treated and each team modeled with the Poisson regression and distribution. Table 3.0 show the model output and observation for each value generated by the model. The attack and defense strength were automatically calculated by the model with respect to the equations stated in the methodology vis-à-vis implementation codes.

TABLE 3: Model information output.

Generalized Linear Model Regression Results			
Dep. Variable:	goals	**No. Observations:**	726
Model:	GLM	**Df Residuals:**	678
Model Family:	Poisson	**Df Model:**	47
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-814.66
Date:	Thur, 14 Nov 2019	**Deviance:**	591.40
Time:	09:33:38	**Pearson chi2:**	528.
No. Iterations:	5	**Covariance Type:**	nonrobust

Figure 5 shows the Poisson model for the number of goal per match between Akwa United and Eyimba International for both the Home and Away match.

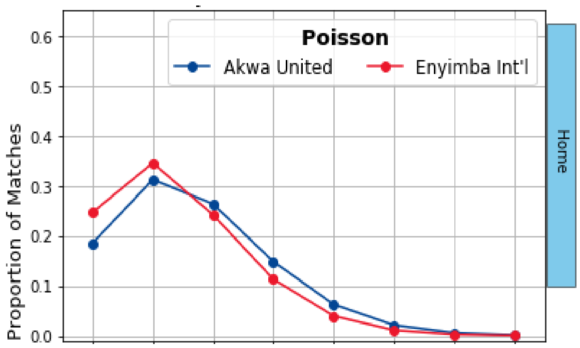


FIGURE 5: Poisson model for two (2) football teams.

4.3 Model Evaluation and Comparison with Bookmarkers Odds

The dataset was shared into training and testing; the testing dataset was set aside for further evaluation of the model. The NPFL super six fixtures were used to carry out the evaluation. The dataset of fixture used and the date the matches were played, the actual win, and the predicted percentage Out of the 15 matches, the model correctly predicted 11 matches correctly which is a percentage of 73.33%. Despite other external factors such as signing of new players, changing of team coach and other social sentiment which was not included in this model. In Table 4.0, the model compared with an online bookmarker's odd. This will help to evaluate the consistency of the model result with the bookmarker's odd and the theory of the mathematical equations. The online bookmarker odds were retrieved from https://hintwise.com/league/Nigeria-NPFL. Table 4.2 shows the comparison between the model and the online bookmarker's odd which was retrieved from www.hintwise.com/league/Nigeria-NPFL as at November 1st 2019.

TABLE 4: Shows the evaluation between the Model Result and the Bookmarker's Odd Using Dataset of NPFL Super Six Fixture 2018/2019.

Srn	Home vs Away Team	Model Predicted Result (%)Home Win(H), Draw(D),Away Win(A)			Bookmarker's Odd in (%) Home Win(H),Draw(D), Away Win(A)			Model Accuracy
1	Enyimba Int'l FC vs Rangers Int'l FC	52.71	35.81	11.49	62.30	25.60	13.10	***True***
2	Kano Pillars Int'l vs Akwa United FC	68.57	20.74	10.69	65.40	24.10	10.50	***True***
3	FC Ifeanyi Ubah vs Lobi Stars FC	49.89	31.45	18.53	54.34	35.23	24.32	***True***
4	Kano Pillars Int'l vs Enyimba Int'l	47.89	35.42	17.17	73.53	28.43	11.65.	***True***
5	Rangers Int'l FC vs Lobi Stars FC	57.93	29.41	12.44	61.00	22.50	16.50	***True***
6	Akwa United FC vs FC Ifeanyi Ubah	64.55	23.12	11.56	59.50	27.60	18.76	***True***
7	Akwa United FC vs Rangers Int'l FC	51.93	29.94	17.94	59.5	27.6	12.9	***True***
8	Lobi Stars FC vs Enyimba FC	43.30	38.43	18.24	59.80	26.41	14.18	***True***
9	Kano Pillars FC vs FC Ifeanyi Ubah	68.27	21.69	9.11	32.4	43.4	23.4.	***True***
10	Enyimba Int'l FC vs FC Ifeanyi Ubah	64.52	27.54	7.65	52.10	26.09	21.00	***True***
11	Akwa United vs Lobi Stars FC	57.23	26.54	15.82	37.00	39.00	24.00	***True***
12	Rangers Int'l vs Kano Pillars Int'l	57.58	28.47	13.69	67.60	22.50	9.90	***True***
13	Lobi Stars FC vs Kano Pillars Int'l	56.44	27.70	15.55	37.00	40.00	23.00	***True***
14	Enyimba Int'l FC vs Akwa United	64.69	25.95	8.96	40.90	31.00	27.70	***True***
15	FC Ifeanyi Ubah vs Rangers Int'l	44.84	34.71	20.40	8.50	27.10	64.4	***False***

Comparing the online bookmaker probabilities with the model output, 14 matches out of 15 were similar which gives 93.33% similarity with other online bookmakers. Using kNN to categorize Home win, Draw, and Away win, the difference in the probabilities of the model and the online book-maker is as a result of some unforeseen contingencies.

5. CONCLUSION

Overall, Football prediction has becoming captivating, as seen in the literatures of different authors from year 1997 to 2021. However, there were difficulties in predicting 100% accurately of outcome of a match especially with the peculiarities and other sentimental factors that come to play in the Nigerian football league. The practical implication of this work, include but not limited to the numerous benefit it pose to make Nigerian football league catch up with their international counterparts, albeit the target audience are football fans and various football club lovers but the predicting the results correctly, larger audience will be reached with this league therefore making it more popular in the international community. The difficulties encountered by specialist in getting hired for international league will also be minimized with the result of this work. Recall, various researchers have used various methods and algorithms to predict the outcome of sport matches. Some algorithms look complex for sport like chess, javelin or even tennis game as it is only one major factor that determines the winner or loser which is the ability of the player. In this research, only team games were considered.

Data mining tool have been maximized for this purpose based on its peculiarities for event prediction. Several predicting models like ML, ANN, BN, LR, SVM, and lazy techniques have been adopted in this work. Several literatures were reviewed vis-à-vis their major drawbacks especially in data availability and evaluation results. This research method allows high prediction accuracy meaning using comprehensive statistics which that gives room for comparison of results with previous studies and available data.

Further study would focus more on sentiment and other factors such as fatigue, change of player, weather conditions, and coaches for a better and reliable model.

6. ACKNOWLEDGEMENT

The authors of this work would like to appreciate NPFL for making available historical data for the public, sport newspapers and magazines among other public sport database used for testing and evaluation purpose in this research.

7. REFERENCES

[1] Adebisi, J. A., Abdulsalam, K. A. and Fawaz O. (2021): IOT Smart Home: Implementation of a real-time Energy Monitoring Pressing Iron. Being a paper International Conference of the Nigeria Computer Society.

[2] Bunker, R. P., and Thabtah, F. (2019). A machine learning framework for sport result prediction. Applied computing and informatics, 15(1), 27-33.

[3] Angelini, G., and De Angelis, L. (2017). PARX model for football match predictions: PARX model for football matches predictions. Journal of Forecasting, 36(7), 795–807.

[4] Baio, G., & Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics, 37(2), 253–264.

[5] Buursma, D. (2010). Predicting sports events from past results. 14th Twente Student Conference on IT, 21. Holland.

[6] Constantinou, A. C., Fenton, N. E., and Neil, M. (2013). Profiting from an inefficient Association Football gambling market: Prediction, Risk and Uncertainty using Bayesian networks. Knowledge-Based Systems, 50, 60–86.

[7] Dixon, M. J., and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), 265–280.

[8] Doyle, P. G., Grinstead, C. M., and Snell, J. L. (2006). Grinstead and Snell's Introduction to Probability.

[9] Esme, E., and Kiran, M. S. (2018). Prediction of Football Match Outcomes Based on Bookmaker Odds by Using k-Nearest Neighbor Algorithm. International Journal of Machine Learning and Computing, 8(1)

[10] Forrest, D., & Simmons, R. (2002). Outcome uncertainty and attendance demand in sport: The case of English soccer. Journal of the Royal Statistical Society: Series D (The Statistician), 51(2), 229–241.

[11] Graham, I., & Stott, H. (2008). Predicting bookmaker odds and efficiency for UK football. Applied Economics, 40(1), 99–109.

[12] Haghighat, M., Rastegari, H., and Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. Advances in Computer Science: An International Journal, 2(5), 7–12.

[13] Haight, F. A. (1967). Handbook of the Poisson distribution.

[14] Home Page—Nigeria Professional Football League. (n.d.). Retrieved October 1, 2019, from https://npfl.ng/

[15] Igiri, C. P., and Nwachukwu, E. O. (2014). An improved prediction system for football a match result. IOSR Journal of Engineering (IOSRJEN), 4(12), 12–20.

[16] Joseph, A., Fenton, N. E., and Neil, M. (2006). Predicting football results using Bayesian nets and other machine learning techniques. Knowledge-Based Systems, 19(7), 544–553.

[17] Karlis, D., & Ntzoufras, I. (2011). Robust fitting of football prediction models. IMA Journal of Management Mathematics, 22(2), 171–182.

[18] Kuhlman, D. (2009). A python book: Beginning python, advanced python, and python exercises. Dave Kuhlman Lutz.

[19] Lessmann, S., Sung, M.-C., and Johnson, J. E. V. (2010). Alternative methods of predicting competitive events: An application in horserace betting markets. International Journal of Forecasting, 26(3), 518–536.

[20] Leung, C. K., and Joseph, K. W. (2014). Sports data mining: Predicting results for the college football games. Procedia Computer Science, 35, 710–719.

[21] Mathworks (n.d.). machinelearning_supervisedunsupervised.png. Retrieved from <https://de.mathworks.com/help/stats/machinelearning_supervisedunsupervised.png>.

[22] Miljkovic, D., Gajic, L., Kovacevic, A., and Konjovic, Z. (2010). The use of data mining for basketball matches outcomes prediction. IEEE 8th International Symposium on Intelligent Systems and Informatics, 309–312.

[23] Müller, A. C., and Guido, S. (2016). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media, Inc.

[24] Murphy, K. P. (2012). Machine learning: A probabilistic perspective. MIT press.

[25] NPFL Data for League 2017. (n.d.). Retrieved from https://npfl.ng/league-table/.

[26] Owramipur, F., Eskandarian, P., & Mozneb, F. S. (2013). Football result prediction with Bayesian network in Spanish League-Barcelona team. International Journal of Computer Theory and Engineering, 5(5), 812.

[27] Passi, K., and Pandey, N. (2018). Increased Prediction Accuracy in the Game of Cricket using Machine Learning. ArXiv Preprint ArXiv:1804.04226.

[28] Purucker, M. C. (1996). Neural network quarterbacking. IEEE Potentials, 15(3), 9–15.

[29] Razali, N., Mustapha, A., Utama, S., and Din, R. (2018). A Review on Football Match Outcome Prediction using Bayesian Networks. Journal of Physics: Conference Series, 1020, 012004.

[30] Reed, D., and O'Donoghue, P. (2005). Development and application of computer-based prediction methods. International Journal of Performance Analysis in Sport, 5(3), 12–28.

[31] Şahin, M., and Uçar, M. (2020). Prediction of sports attendance: A comparative analysis. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, 1754337120983135.

[32] Sharon Andrews & Mark Sheppard (2020):Software Architecture Erosion: Impacts, Causes, and Management. International Journal of Computer Science and Security (IJCSS), Volume (14) : Issue (2) : 2020

[33] Tavares, A. T. (2018). Predicting Results of Brazilian Soccer League Matches. University of Wisconsin-Madison.

[34] Tina T. (2020): Social Big Data: Techniques and Recent Applications. International Journal of Computer Science and Security (IJCSS), Volume (14): Issue(5): 2020.

[35] Ujwal, U. J., Antony, P. J., and Sachin, D. N. (2018). Predictive Analysis of Sports Data using Google Prediction API. International Journal of Applied Engineering Research, 13(5), 2814–2816.

[36] Van Eetvelde, H., Mendonça, L. D., Ley, C., Seil, R., and Tischer, T. (2021). Machine learning methods in sport injury prediction and prevention: a systematic review. Journal of experimental orthopaedics, 8(1), 1-15.

[37] Wilkens, S. (2020). Sports prediction and betting models in the machine learning age: The case of tennis. Journal of Sports Analytics, (Preprint), 1-19.

[38] Yang, S., Luo, L., and Tan, B. (2021). Research on Sports Performance Prediction Based on BP Neural Network. Mobile Information Systems, 2021.

[39] Zaveri, N., Tiwari, S., Shinde, P., Shah, U., and Teli, L. K. (2018). Prediction of Football Match Score and Decision Making Process. International Journal on Recent and Innovation Trends in Computing and Communication, 6(2), 162–165.

[40] Zdravevski, E., and Kulakov, A. (2009). System for Prediction of the Winner in a Sports Game. International Conference on ICT Innovations, 55–63. Springer.

A Novel Approach To The Weight and Balance Calculation for The De Haviland Canada DHC-6 Seaplane Operators

Houssam Hammoudi *houssam@houssamhammoudi.com*
CEO, TradeSec Corp
Calgary, T3N 1T4, Canada

Abstract

The main objective of this research is to provide companies operating different fleets of the De Havilland Canada Twin Otter DHC-6 seaplanes with an alternative method to the time-consuming Whizz Wheel procedure when calculating the weight and balance. Using this application, these operators can lower their aircraft turnaround, speed up the passenger boarding, dispatch the flights efficiently and save on fuel and dock expenses. Furthermore, this research shows how operators do their calculations currently and the positive impact of the application on their entire operation, including extra revenue generation amounting to $4M per year. Most DHC-6 seaplane operators are mainly in the Maldives. Therefore, this research was conducted while piloting these seaplanes and studying the day-to-day operations. While this paper presents the implementation of this software and its design model, it also discusses how two major operators used this application in the Maldives and one in St Vincent and the Grenadines.

Keywords: Software Engineering, Android, Java, Mobile, Seaplanes, DHC-6, W&B.

1. INTRODUCTION

Aviation is an essential lifeline of the world’s economy; it supports more than 65.5 million jobs worldwide and enables $2.7 trillion in global GDP (Shewring & Stevens, 2010). The aviation sector, particularly in the Maldives, makes a significant contribution to the country’s economy. The Maldives alone supports 73,000 jobs and a $3.1 billion added contribution to the GDP. In addition, air transport and tourists arriving by air represent 58.8% of the Maldives GDP. There are five major airlines in the Maldives, and two of these airlines operate the most extensive fleet of DHC-6 seaplanes in the world, conducting 24,700 flights, 51,700 takeoffs, and landings each year. With only 14 airports, the Maldives relies heavily on their seaplane fleets to link their scattered atolls (Air Transport Action Group, 2021). With this high frequency of flights, the Maldivian airlines need to optimize their weight and balance calculations procedures to speed up each aircraft’s turnaround and dispatch their fleet quicker to their destinations. Saving time on the dispatch process means that the company can conduct more flights without increasing its crew number or fleet size.

In this research, we present the overall approach we chose to use, the method we used to calculate the weight and balance accurately. Finally, we give you the results of our optimization, quantifying the time and money saved per flight.

2. OBJECTIVES

The study aims to develop an Android-based weight and balance and aircraft performance calculation app replacing the traditional whizz wheel to save time and money for seaplane operators worldwide.

3. LITERATURE REVIEW

While reviewing the literature on similar research papers, we encountered that most research focuses on transport category narrow and wide-body passenger jets with 20 seats and up. We, however, decided to focus on the commuter category aircrafts with a maximum of 19 seats and down. These commuter aircrafts are used by small operators who benefit the most from this software implementation. This niche market is very small compared to the mainstream aviation industry. The profits from this market are very limited and this is why there is less interest from software companies or manufacturers to tackle this problem.

3.1 The Importance of The Weight and Balance

Aircraft accidents related to weight and balance issues are yearly occurrences. These accidents occur due to incorrect loading and or use of wrong weight for takeoff performance calculations (Cokorilo et al., 2010). Aircraft manufacturers offer automated systems to calculate weight and balance; however, in the case of the DHC-6 among other planes, the “whizz wheel” is the only way to perform these calculations (Figure 1).

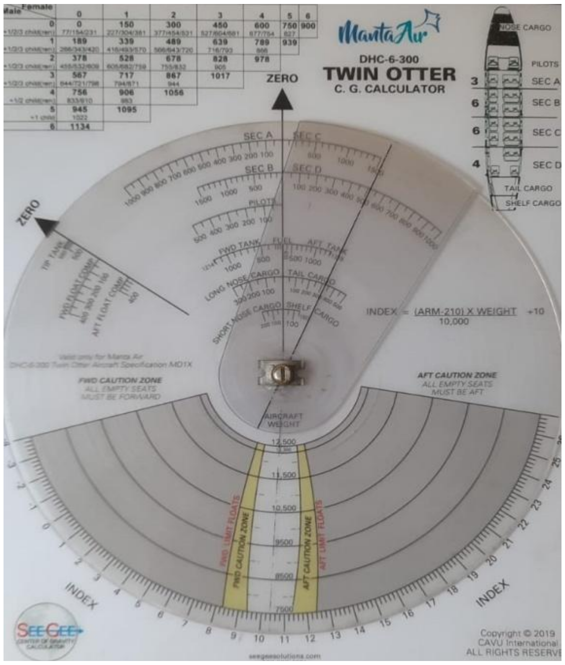


FIGURE 1: C. G. Whizz Wheel calculator for the DHC-6 aircraft.

The whizz wheel is a navigation computer that can perform many calculations and have many variables as inputs such as the wind direction, wind speed, load factors, etc.

The pilot turns the wheel either right or left, aligns the wheel with different indexes, etc. Although they can do many calculations, these whizz wheels are time-consuming; it can take up to 3 minutes to finalize a weight and balance of a fully-loaded DHC-6 seaplane. Also, any change to the loading, either a passenger moved to another seat or extra suitcases loaded or unloaded, last-minute arrivals, the calculation needs to be re-done from scratch. This problem is precisely what we want to fix.

3.2 Navigation and Positioning Calculation Restrictions

The Maldivians operate these aircraft in the most remote areas in the world; the application cannot rely on assisted GPS from android OS but rather use mathematics to calculate many aspects:

- The Azimuth from the latitude and longitude of the seaplane base
- The proper heading to the destination seaplane base
- The distance and fuel consumption depicted as sector burn

3.3 Weight and Balance Calculation

The Mean Aerodynamic Chord, also known as the "Percent Mac" calculation, gives us a value of the %Mac at takeoff and landing. Having this information, the pilot knows whether the plane will be within its center of gravity or end up with an aft or forward center of gravity at any phase of the flight. Finally, we chose the average weight of passengers based on their gender and age, Female "F," Male "M," and Child "C."

4. PROJECT DESIGN AND METHODOLOGY

The methodology used in this study follows the "prototyping model" software development Methodology. This model allows the software prototype to be built, tested, and refined on end-user feedback until a good working version is accomplished.

4.1 The Requirements Gathering and Analysis Phase

The requirement gathering and analysis phase consisted of being on the field gathering all the observations made by the pilots, the dispatchers, and the company management. In addition, during the study, we witnessed the problems caused by the current process, and we had the opportunity to confirm our assumptions.

4.2 The Quick Design Phase

In this section, we produced a quick and straightforward framework showing how the user interface will look.

4.3 The Build Phase

This section produced a quick and simple design that we call an MVP a "Minimum Viable Product." This version of the software included only four or five features. This fast design was then presented to the end-user to gather more feedback integrated into future design phases.

4.4 Field Evaluation Phase

In this section, we used the application in the field and gathered valuable feedback from the end-user, allowing us to determine its strengths, weaknesses, and must-have features.

4.5 The Prototype Phase

In this section, we measured the impact of the application, we determined the nice to have features and the must-have features, and we were able to have an approved final prototype.

5. THE USER INTERFACE & NAVIGATION

We chose to develop the application on Android OS because it provides a huge selection of pre-built UI components like UI controls and structured layouts, allowing us to build a graphical user interface quickly. We also chose Android because it is based on Java which is a powerful object-oriented programming language.

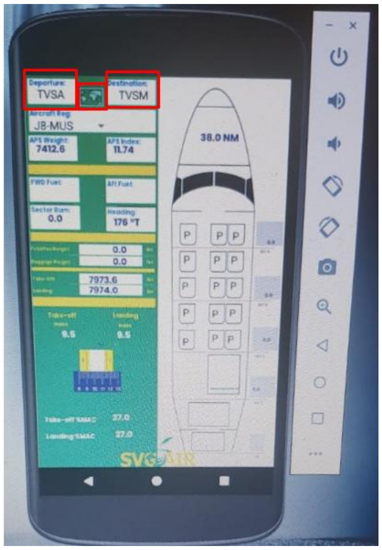
5.1 Departure and Destination Selection

5.1.1 Departure and Destination UI Component

This UI section allows the pilot to select the departure and destination airport from the airport database. Each airport has the following attributes:

- Airport ICAO code
- Airport name
- Airport longitude
- Airport latitude

Once the departure and destination airports are chosen, we can launch the backend distance calculation by pressing the Globe button depicted in figure 2.



Departure: TVSA

Destination: TVSM

Aircraft Reg: J8-MUS

APS Weight: 7412.6

APS Index: 11.74

FWD Fuel:

Aft Fuel:

Sector Burn: 0.0

Heading: 176 °T

0.0

0.0

7973.6

7974.0

Take-off 9.5

Landing 9.5

Take-off %MAC 27.0

Landing %MAC 27.0

FIGURE 2: The departure and destination selection.

5.1.2 Backend Distance Calculation

The backend component is the code that runs to calculate the distance between two locations. Figure 3 shows this backend code using the Haversine formula to accomplish the calculation. This calculation is crucial to calculate and display the fuel burn to the pilot as shown in figure 4.

Formula to calculate the distance between two points on earth :

$$a = \sin^2(\Delta\varphi / 2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda / 2)$$

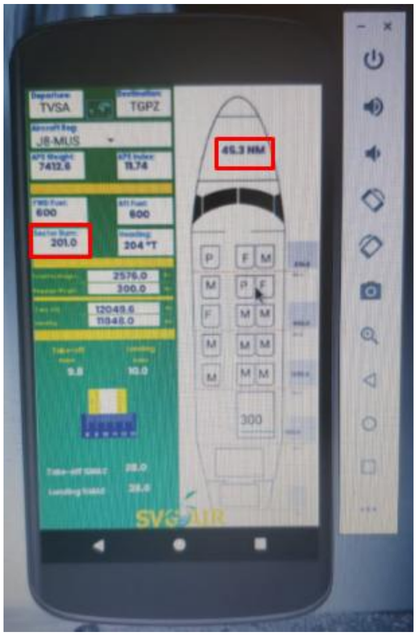
$$c = 2 \cdot \operatorname{atan2}\left[\sqrt{a},\ \sqrt{(1-a)}\right]$$

$$d = R \cdot c$$

Where :

φ_1 *is the latitude of initial point (positive for N and negative for S)*
φ_2 *is the latitude of the final point (positive for N and negative for S)*
λ_1 *is the longitude of the initial point (positive for E and negative for W)*
λ_2 *is the longitude of the final point (positive for E and negative for W)*
$\Delta\varphi = \varphi_2 - \varphi_1$
$\Delta\lambda = \lambda_2 - \lambda_1$
R is the radius of the Earth in meters ($R = 6371000\ m$)

FIGURE 3: The haversine formula to calculate the distance between two locations.



Departure: TVSA
Destination: TGPZ
Aircraft Reg: J8-MUS
APS Weight: 7412.6
APS Index: 11.74
FWD Fuel: 600
Aft Fuel: 600
Sector Burn: 201.0
Heading: 204 °T
2576.0
300.0
12049.6
11048.0
9.8
10.0
Take-off 28.0
Landing 28.0
45.3 NM

FIGURE 4: Sector Burn for the flight in pounds as well as the distance in Nautical Miles.

5.1.3 Backend Azimuth Calculation

Now that we have the distance between two points, we need to calculate the Azimuth. To do so, we use the same latitude and longitude of the departure and destination airport and input them into the azimuth formula below.

$$\textit{Formula to calculate the Azimuth between two sets of coordinates:}$$

$$\theta = \text{atan2}\,[(\sin\Delta\lambda \cdot \cos\varphi_2),\ (\cos\varphi_1 \cdot \sin\varphi_2 - \sin\varphi_1 \cdot \cos\varphi_2 \cdot \cos\Delta\lambda)]$$

FIGURE 5: The haversine formula to calculate the Azimuth between two locations.

5.1.4 Backend Bearing Degree Calculation

Before using the azimuth radian, we need to convert it to a positive degree bearing that the pilots can use as their flight heading as illustrated in figure 6.

$$\textit{Formula to convert radians to degrees:}$$

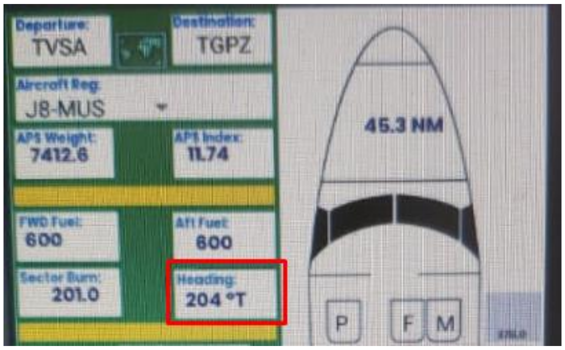
$$\theta = \text{rad} \times 180° \div \pi = \text{Bearing degree}$$

FIGURE 6: The formula to convert radians to degrees.

If the resulting bearing degree is negative, we must convert it to a positive bearing degree. To do that, we add 360 to the negative result to get a positive bearing degree as illustrated in figure 7. The bearing we also call the true heading, in this case, is displayed in the heading box of the UI as depicted in figure 8.

$$\textit{Converting negative bearing degree to positive bearing degree:}$$
$$\text{negative bearing degree} + 360 = \text{positive bearing degree}$$

FIGURE 7: Convert negative bearing degree to positive bearing degree.



Departure: TVSA

Destination: TGPZ

Aircraft Reg: J8-MUS

APS Weight: 7412.6

APS Index: 11.74

FWD Fuel: 600

Aft Fuel: 600

Sector Burn: 201.0

Heading: 204 °T

FIGURE 8: The true heading displayed for the pilots.

5.2 Aircraft Registration Selection

5.2.1 Aircraft Registration UI Component

This section of the UI allows the pilot to select the aircraft that will conduct the flight. Although it can be the same model of aircraft, each plane has its subtle differences. For example, three airplanes of the same model can have a different APS weight, Basic weight, or Horizontal arm. The aircraft numbers we will be using are:

- APS Weight or OEW weight
- Horizontal Arm
- Center of Gravity
- Moment

The **APS "Aircraft Prepared for Service"** and the **OEW "Operating Empty Weight**" describe an empty aircraft with non-useable fuel, engine oil, and equipment required to conduct the flight, excluding any useable fuel cargo and people.

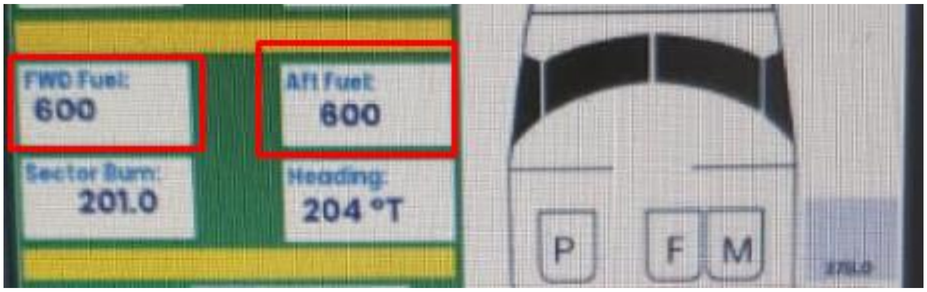
The Horizontal ARM is the distance from the reference datum to the center of gravity of an item.

The center of gravity CG is the point at which the weight of a body is assumed to be concentrated. In other words, for an aircraft is the point over which the aircraft would maintain balance. If the weight is focused forward of the CG, we call this a Forward C of G, meaning the aircraft will pitch down while an Aft C of G means that the weight is concentrated Aft of the CG. The CG must stay within limits throughout all the phases of flight for the aircraft to maintain its flight characteristics.

The Moment is the force that causes an object to rotate, and it is used to calculate the center of gravity.

5.3 Fuel Quantity Input

This UI section allows the pilot to enter how much fuel is in the forward and the aft fuel tanks. Each tank has its ARM, which is the distance from the reference datum. The fuel weight in pounds is multiplied by its corresponding arm depending on which tank it resides in and gives us a Moment. Thus, the sum of the fuel weight in the two tanks is instantaneously added to the total aircraft weight as shown below in figure 9.



FWD Fuel: 600

Aft Fuel: 600

Sector Burn: 201.0

Heading: 204 °T

FIGURE 9: Fuel quantity input in forward and aft tank.

5.4 Passenger Loading

This section of the UI allows the pilot to set up the passenger loading and move passengers around, allowing the aircraft to stay within its center of gravity limit at all phases of the flight. This is an advantage since the calculations are made instantaneously while the pilot is setting up the loading. In this section, each seat has its corresponding ARM. To simplify the math, we assumed the average weight of females, males, and children as per ICAO regulations. Therefore, each seat can be one of the following states:

- **P** as in “Passenger”: The seat is vacant.
- **M** as in “Male”: The seat is occupied by a male.
- **F** as in “Female”: The seat is occupied by a female.
- **C** as in “Child”: The seat is occupied by a child.

Each row has its sector weight as well. The importance of each sector has its arm and will be added to the total weight and the C of G calculation as shown in the figure below.

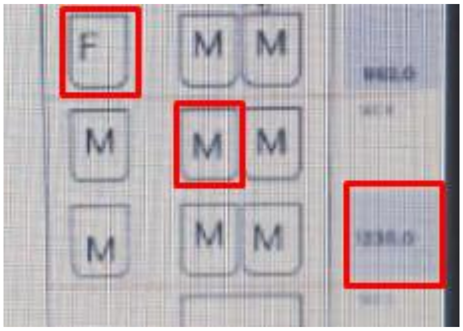
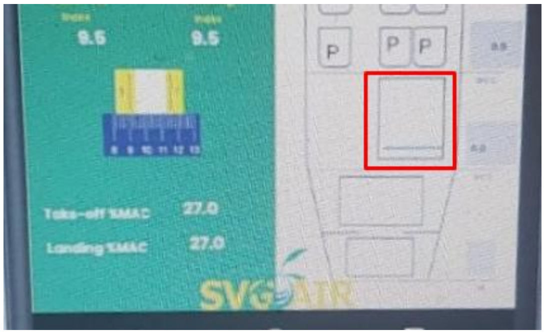


FIGURE 10: Passenger loading and sector weight.

5.5 Luggage Loading

This section of the UI allows the pilot to input the luggage weight in pounds in the luggage area at the back of the aircraft. Like any other position in the aircraft cabin, the luggage area also has its ARM value. Therefore, the luggage weight will be multiplied by its corresponding arm to give us a Moment to calculate the index and be added to the total aircraft weight as shown in the figure below.



Index 9.5 Index 9.5

Take-off %MAC 27.0

Landing %MAC 27.0

FIGURE 11: Luggage Loading and sector weight.

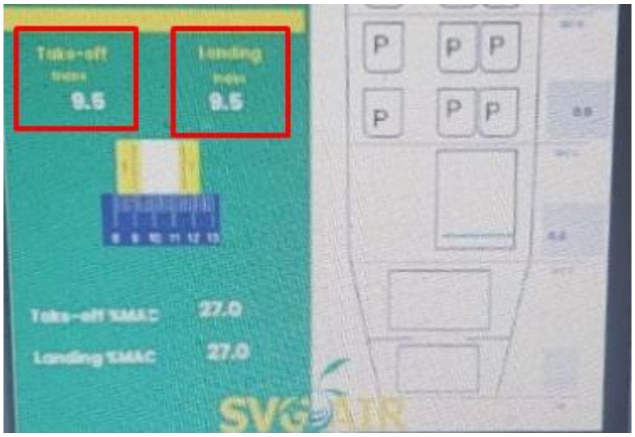
5.6 Aircraft Index

The aircraft index is the Moment divided by a constant in our case, 10.000 as per the aircraft operating manual. We use the index to simplify computations of aircrafts' long arms, resulting in a sizeable unmanageable number. We use the index formula illustrated in figure 12 and display the result in the takeoff index box as shown in figure 13.

$$\textit{Formula to find the index of an aircraft:}$$

$$\text{Index} = 10 + \frac{\text{Basic Weight (Horizontal Arm} - 210)}{10000}$$

FIGURE 12: Formula to calculate the aircraft's index.



Take-off Index 9.5

Landing Index 9.5

Take-off %MAC 27.0

Landing %MAC 27.0

FIGURE 13: Index on takeoff and landing.

5.6.1 Aircraft Index Landing

The landing index is calculated by factoring in the only variable changing with time, fuel consumption, and re-doing the same index calculation.

5.7 %MAC Calculation

The %Mac "Percent Mean Aerodynamic Chord" is a calculation showing the center of gravity over the wing. It is mandatory to have the aircraft %MAC within the limit during all flight, takeoff, cruise, and landing phases.

5.7.1 Takeoff %MAC

The formula to calculate the %MAC in figure 14 involves three different parameters:

- LEMAC: The inches aft of the datum of the leading edge of the MAC
- ARM: The distance from the reference datum to the center of gravity of an item
- MAC: The actual length of the MAC

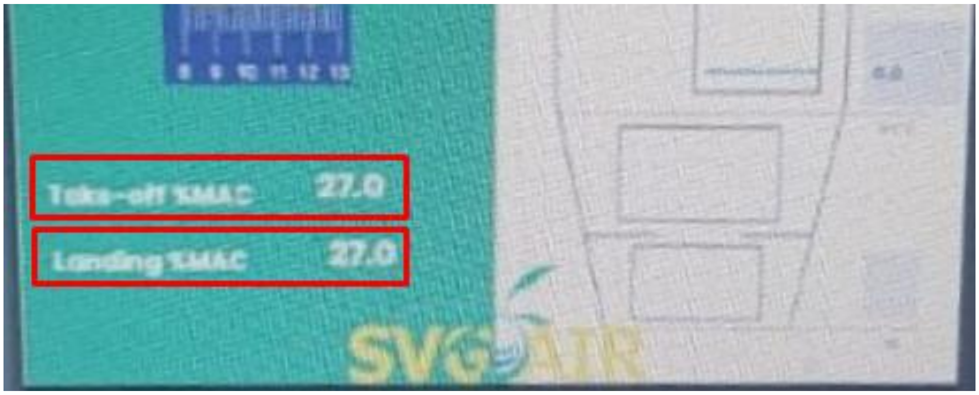
Formula to calculate the %MAC of an aircraft:

$$\%\mathrm{MAC} = \frac{\mathrm{ARM} - \mathrm{LEMAC} \times 100}{\mathrm{MAC}}$$

FIGURE 14: Index on takeoff and landing.

5.7.2 Landing %MAC

The formula to calculate the landing %MAC is the same. The only difference is that we account for the fuel consumption rate and multiply it by flight time. Both takeoff and landing %MAC are shown to the pilots on the UI as you can see on figure 15.



Take-off %MAC 27.0

Landing %MAC 27.0

FIGURE 15: Index on takeoff and landing.

6. CONCLUSIONS

The developed application helped two operators speed up their dispatch rate. Before having this solution, the operators spend 3 minutes using the whizz wheel, assuming no errors are made. The process took even longer on average due to last-minute unloading, last-minute passenger boarding, and extra luggage loaded. This solution allowed them to finish all calculations in under 1 minute and without errors. To quantify the financial benefits of using this app, we can multiply the 2 minutes saved per flight multiplied by 50, which is the number of flights per day, giving us 1.61 hours saved each day. These savings equate to 1 full revenue-generating flight, meaning the airline can conduct their current flights without extra planes and add one additional flight to their schedule. Each hour of a revenue flight generates approximately 7000$ dollars. We then can assume that the company can generate an extra 11,316$ every day, which amounts to $4M of revenue per year. Although we could quantify the financial benefits of adopting this application, the main unquantifiable benefits are the convenience for the flight crew that has less workload and, most importantly, the enhanced safety of the flights.

7. REFERENCES

Air Transport Action Group. (2021). *Social and economic benefits of aviation*. Retrieved August 10, 2021, from https://www.atag.org/our-activities/social-and-economic-benefits-of-aviation.html/.

Air Transport Action Group. (2020, September). Aviation: *Benefits beyond borders.* Aviation Benefits Beyond Borders. https://aviationbenefits.org/downloads/aviation-benefits-beyond-borders-2020/.

Brindisi, A., Ameduri, S., Concilio, A., Ciminello, M., Leone, M., Iele, A., Consales, M., & Cusano, A. (2019). A multi-scaled demonstrator for aircraft weight and balance measurements based on FBG sensors: Design rationale and experimental characterization. *Measurement*, 141, 113-123. https://doi.org/10.1016/j.measurement.2019.03.014.

Bogna, S.(2021, April 09).*How to calculate the azimuth from latitude and longitude.*Omni calculator.https://www.omnicalculator.com/other/azimuth#how-to-calculate-the-azimuth-from-latitude-and-longitude

Čokorilo, O., Gvozdenović, S., Vasov, L., & Mirosavljević, P. (2010). Analysis of aircraft weight and balance related safety occurrences. *Journal of Applied Engineering Science*, *8*(2), 83-92.

Flight Literacy. *Center of Gravity (CG) and Mean Aerodynamic Chord (MAC)*. Retrieved July 13, 2021, from https://www.flightliteracy.com/center-of-gravity-cg-and-mean-aerodynamic-chord-mac/.

Shewring, P., & Stevens, R.(2010).*Canadian airline transport pilot licence workbook*: *a comprehensive guide to prepare pilots to write the ATPL examination*. Aero Course.

Souffriau, W., Vanden, B., Greet.(2011).*A Mixed Integer Programming Approach to the Aircraft Weight and Balance Problem.* Procedia - Social and Behavioral Sciences, 20(2011), 1051-1059. https://doi.org/10.1016/j.sbspro.2011.08.114.

A Cost-effective Automated Weather Reporting System AWRS for The Canadian Remote Northern Air Operators

Houssam Hammoudi *houssam@houssamhammoudi.com*
CEO, TradeSec Corp
Calgary, T3N 1T4, Canada

Abstract

Air transportation is essential for Canada's and US northern communities. It is the leading lifeline supplying fresh food, medicine, and other goods; providing Health care services; medical emergency evacuation; and supporting travel outside of the communities. In addition, air transportation is the only reliable year-round mode of transportation. However, the Canadian north and Alaska present significant operational challenges mainly due to inhospitable terrain, harsh conditions, extreme cold. The challenges are financial as well due to low passenger volumes and high operations costs. This research shows how northern Air operators can enhance flight safety using WX-Ready as an AWRS "Automated Weather Information System" to get vital weather information without investing in expensive commercially available ACARS systems (Government of Canada, 2017).

Keywords: Software Engineering, ACARS, Python, Linux, METAR, TAF, Aviation, Air Operators.

1. INTRODUCTION

Aviation is an essential lifeline of the world's economy; it supports more than 65.5 million jobs worldwide and enables $2.7 trillion in global GDP (Government of Canada, 2017). The aviation sector in Canada generates more than 632000 jobs and $49 billion to contribute to the Canadian GDP. Although Canada has 500 airports, only about half of these airports have scheduled commercial flights. However, most of these airports are small. From these 500 airports, only 117 are remote northern airports scattered across the vast Canadian north. In addition, many northern communities operate ice strips in the arctic region, which are not certified nor listed as airports by Transport Canada. (Government of Canada, 2017).

2. OBJECTIVES

The study aims to develop a Python/Linux Weather retrieval and display system that replaces the expensive commercially available ACARS systems and enhances flight safety without a significant investment. Small operators cannot afford a half-million dollars ACARS solution. Our objective is to fill this need and fix this operational problem in the industry by providing an affordable AWRS system that can enhance flight safety at minimal costs. In addition, we will leverage the GSM/Cellular network and its SMS capability to send and retrieve messages. We will also set a concept for future development using Iridium Satellite Short Burst Data SBD messaging.

3. LITERATURE REVIEW

Reviewing similar research papers, we found that most of the research papers focused on operations within commercial airports (Benjamin & Moninger, 2016). We, however, focused on northern operators operating in and from ice strips and small northern airports. We realized that most of the literature is only aimed at established airlines with big commercial jets; we couldn't find anything related to small operators, charter or commuter operators with aircrafts less than 19 passengers. Another observation was that none of the studies covered Medivac operators that are vital in Canada (International Civil Aviation Organization, 2010).

3.1 The Importance of Aviation Weather

The aviation weather systems have evolved tremendously since the 1980s; in addition to fixed weather observation stations, aircraft with their onboard sensors take atmospheric measurements and share that information with other aircraft and ground stations via the Aircraft Meteorological Data Relay (AMDAR). (Anaman et al., 2017; Federal Aviation Administration, 2009).

3.2 Important Weather Readings

Aviation weather information comes in the form of forecasts (Federal Aviation Administration, 2009) observations, and notices. We are going to focus only on the following:

- **METAR** – Meteorological Terminal Air Report
- **TAF** – Terminal Aerodrome Forecast
- **NOTAM** – Notice to airmen

4. PROJECT DESIGN AND METHODOLOGY

The methodology used in this study follows the software prototyping development model (Susanto et al., 2019) This model allows building, testing, and refining the software prototype based on end-user feedback until a good working version is accomplished.

4.1 The Requirements Gathering and Analysis Phase

The requirement gathering and analysis phase consisted of being on the field gathering all the observations made by the pilots, the dispatchers, and the company management. We witnessed the lack of weather information during the study during ferry flights over remote northern areas and transoceanic flights.

4.2 The Quick Design Phase

This section presents a quick and straightforward framework showing how the user interface will appear.

4.3 The Build Phase

This section produces a quick and simple design that we call an MVP a “Minimum Viable Product.” This version of the apparatus included only four features. This phase allowed us to gather valuable feedback about the product

4.4 Field Evaluation Phase

In this section, we describe how we used the apparatus in the field during transoceanic flights and flights conducted in the Arctic region. This phase was crucial to gathering valuable feedback from the pilots, allowing us to determine its strengths and weaknesses and the must-have features.

4.5 Refining The Prototype Phase

In this section, we measured the impact of the apparatus, we determined the nice-to-have features and the must-have features, and we were able to have an approved final prototype.

5. HARDWARE COMPONENTS

This section will describe in detail the different hardware components used in this project and explain their roles.

5.1 Single Board Computer

This apparatus combines software and a single board computer that needs to bring a small size factor, versatility, and upgrade either by expansion boards or hardware attached on top boards (HAT) and cost-effective. We found all these characteristics in the Raspberry Pi 3 Model B+ shown in figure 1 below.

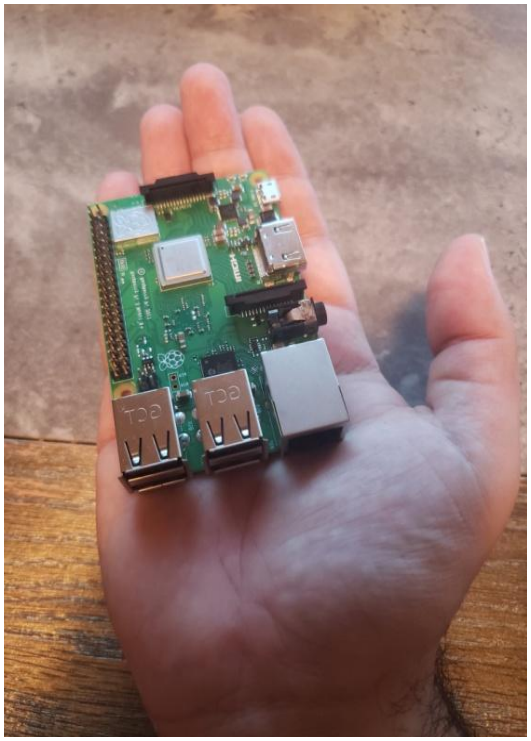


FIGURE 1: Raspberry Pi Model B+.

5.2 SIM800 GSM/GPRS HAT Expansion Board

This HAT board is explicitly made for the Raspberry Pi. It attaches on top of the Pi by connecting to the 40 GPIO pins shown in figure 2.

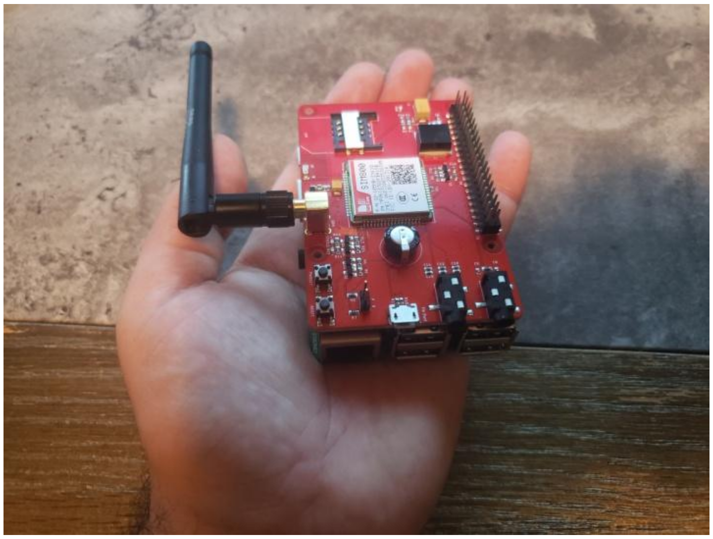


FIGURE 2: SIM800 HAT board connected to the Raspberry Pi.

5.3 TFT 3.5-inch LCD Touch Screen

The 3.5-inch LCD touch screen displays the weather data and allows users to make selections and input airport codes. The LCD screen connects to the raspberry pi via the 40 GPIO pins of the SIM800 board shown in figure 3.

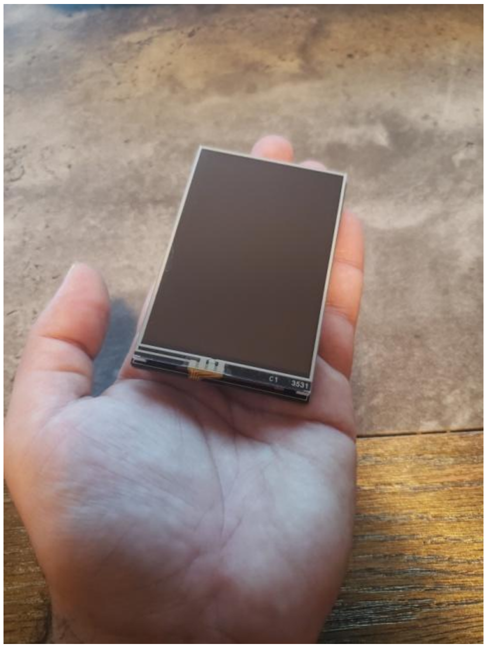


FIGURE 3: TFT 3.5-inch LCD Touch Screen.

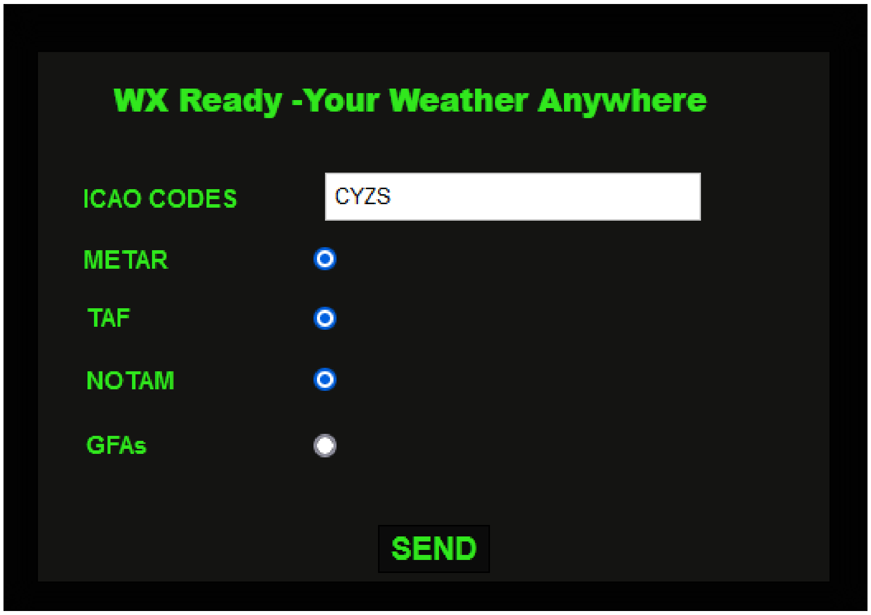
6. USER INTERFACE AND NAVIGATION

We chose to develop the Linux Debian-like OS called Raspbean because it provides a cost-effective, stable, and robust platform. We also decided to write the UI using the Tkinter Python framework due to its various GUI elements and ease of use.

6.1 Airport Selection

6.1.1 ICAO CODES UI Component

This section of the UI allows the pilot to enter all airports they wish to retrieve weather info. Commas can separate the airport codes. The crew members can select which information they want to receive by selecting the METAR, TAF, NOTAM, and GFAs as depicted in figure 4.



WX Ready -Your Weather Anywhere

ICAO CODES: CYZS

METAR ●

TAF ●

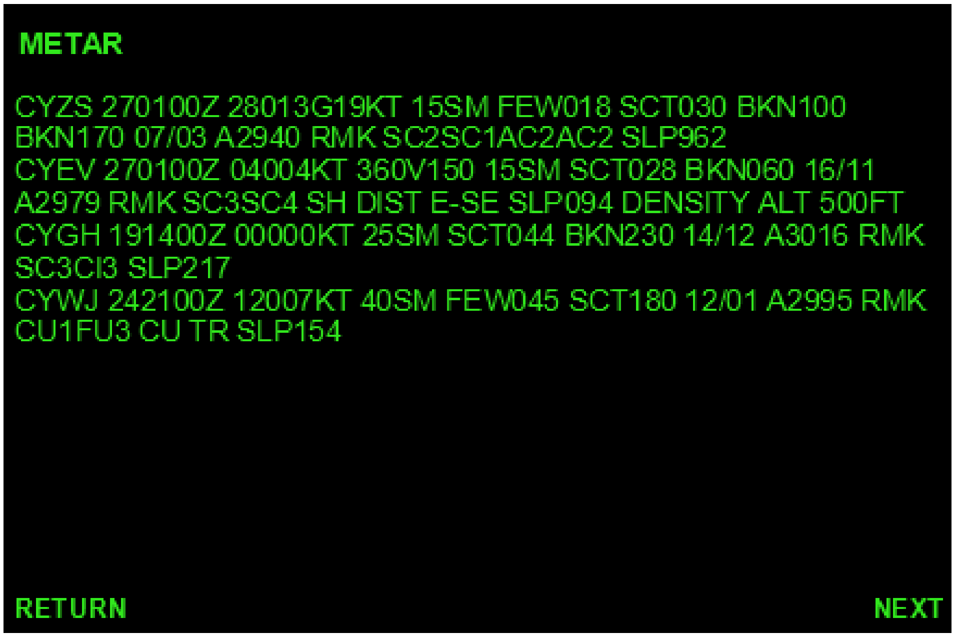
NOTAM ●

GFAs ○

FIGURE 4: The airport, METAR, TAF, and Notam selection.

6.1.2 METAR Weather Info Display UI Component

This screen depicted in figure 5 below shows the METAR information of every airport in the airport list.

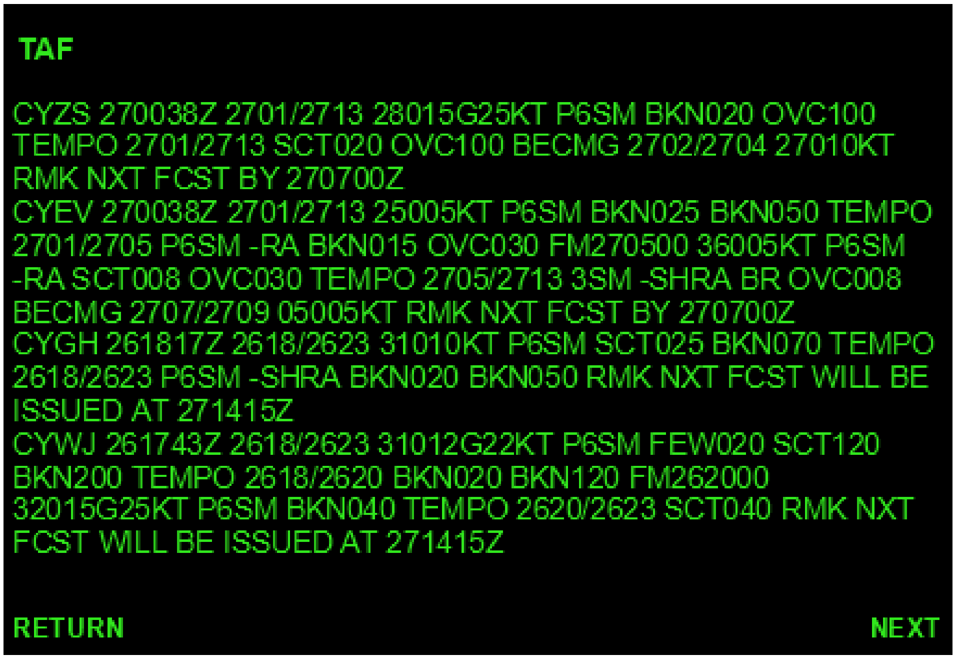


```
CYZS 270100Z 28013G19KT 15SM FEW018 SCT030 BKN100
BKN170 07/03 A2940 RMK SC2SC1AC2AC2 SLP962
CYEV 270100Z 04004KT 360V150 15SM SCT028 BKN060 16/11
A2979 RMK SC3SC4 SH DIST E-SE SLP094 DENSITY ALT 500FT
CYGH 191400Z 00000KT 25SM SCT044 BKN230 14/12 A3016 RMK
SC3CI3 SLP217
CYWJ 242100Z 12007KT 40SM FEW045 SCT180 12/01 A2995 RMK
CU1FU3 CU TR SLP154
```

FIGURE 5: METAR information displayed.

6.1.3 TAF Weather Info Display UI Component

The figure 6 below shows the METAR information of every airport in the airport list.

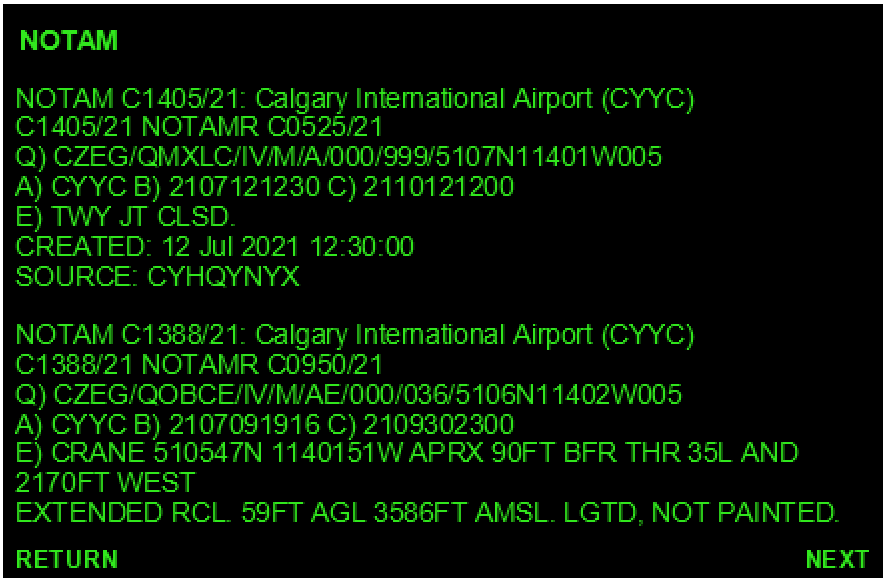


```
CYZS 270038Z 2701/2713 28015G25KT P6SM BKN020 OVC100
TEMPO 2701/2713 SCT020 OVC100 BECMG 2702/2704 27010KT
RMK NXT FCST BY 270700Z
CYEV 270038Z 2701/2713 25005KT P6SM BKN025 BKN050 TEMPO
2701/2705 P6SM -RA BKN015 OVC030 FM270500 36005KT P6SM
-RA SCT008 OVC030 TEMPO 2705/2713 3SM -SHRA BR OVC008
BECMG 2707/2709 05005KT RMK NXT FCST BY 270700Z
CYGH 261817Z 2618/2623 31010KT P6SM SCT025 BKN070 TEMPO
2618/2623 P6SM -SHRA BKN020 BKN050 RMK NXT FCST WILL BE
ISSUED AT 271415Z
CYWJ 261743Z 2618/2623 31012G22KT P6SM FEW020 SCT120
BKN200 TEMPO 2618/2620 BKN020 BKN120 FM262000
32015G25KT P6SM BKN040 TEMPO 2620/2623 SCT040 RMK NXT
FCST WILL BE ISSUED AT 271415Z
```

FIGURE 6: TAF information displayed.

6.1.4 NOTAM Weather Info Display UI Component

The figure 7 below shows the NOTAM information of every airport in the airport list.



```
NOTAM

NOTAM C1405/21: Calgary International Airport (CYYC)
C1405/21 NOTAMR C0525/21
Q) CZEG/QMXLC/IV/M/A/000/999/5107N11401W005
A) CYYC B) 2107121230 C) 2110121200
E) TWY JT CLSD.
CREATED: 12 Jul 2021 12:30:00
SOURCE: CYHQYNYX

NOTAM C1388/21: Calgary International Airport (CYYC)
C1388/21 NOTAMR C0950/21
Q) CZEG/QOBCE/IV/M/AE/000/036/5106N11402W005
A) CYYC B) 2107091916 C) 2109302300
E) CRANE 510547N 1140151W APRX 90FT BFR THR 35L AND
2170FT WEST
EXTENDED RCL. 59FT AGL 3586FT AMSL. LGTD, NOT PAINTED.
RETURN                                                NEXT
```

FIGURE 7: NOTAM information displayed.

7. SYSTEM ARCHITECTURE

The system architecture depicted in figure 8 is comprised of different components and services.

- Webserver containing the web service listener
- SMS 3rd party service API "Twilio."
- SMS 3rd party webhook service "Twilio."
- SIM800 receiving and sending SMS messages

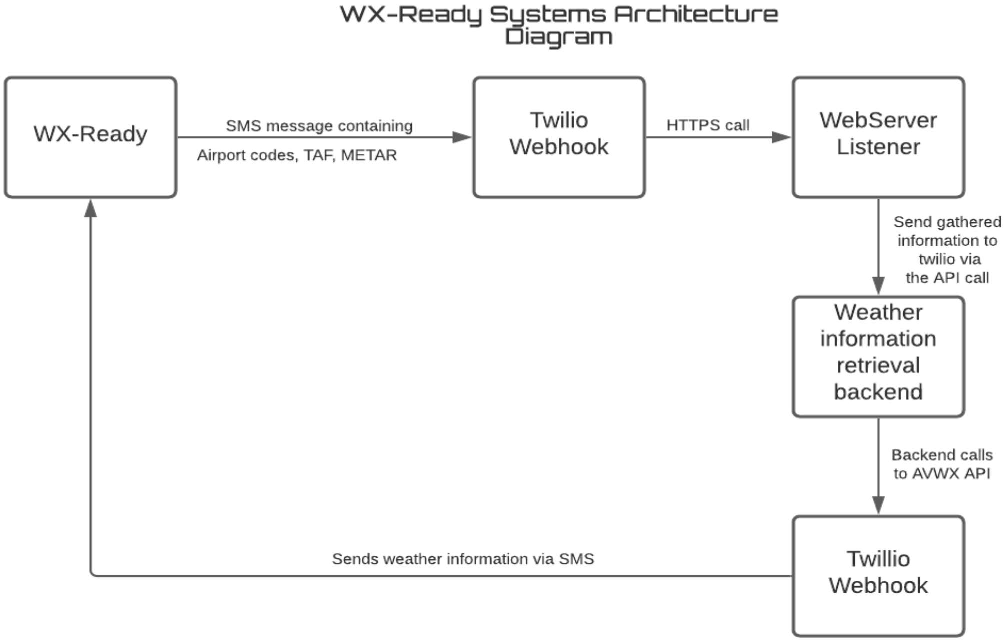

WX-Ready Systems Architecture Diagram
WX-Ready
SMS message containing
Airport codes, TAF, METAR
Twilio Webhook
HTTPS call
WebServer Listener
Send gathered information to twilio via the API call
Weather information retrieval backend
Backend calls to AVWX API
Twillio Webhook
Sends weather information via SMS

FIGURE 8: Systems architecture diagram.

7.1 Webserver

The web server is an Apache server running on a Debian Linux OS. We chose Apache for its ease of use and stability. The webserver is hosted publicly and responds only to HTTPS calls for added security.

7.2 Twilio Messaging Service

Twilio is a cloud communication platform allowing developers to send calls, SMS, and MMS messages using their many web service APIs. Moreover, the inbound and outbound SMS message prices are minimal, costing only $0.0075 per message. This price point makes the WX-Ready very attractive and advantageous for northern operators who cannot afford a hundred thousand dollars ACARS system.

7.3 Twilio Webhook Service

Twilio is a cloud communication platform allowing developers to send calls, SMS, and MMS messages using their many web service API’s.

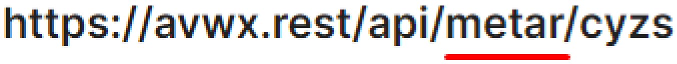
7.4 Twilio Webhook Service

Twilio is a cloud communication platform allowing developers to send calls, SMS, and MMS messages using their many web service API’s.

8. WEATHER INFORMATION RETRIEVAL BACKEND

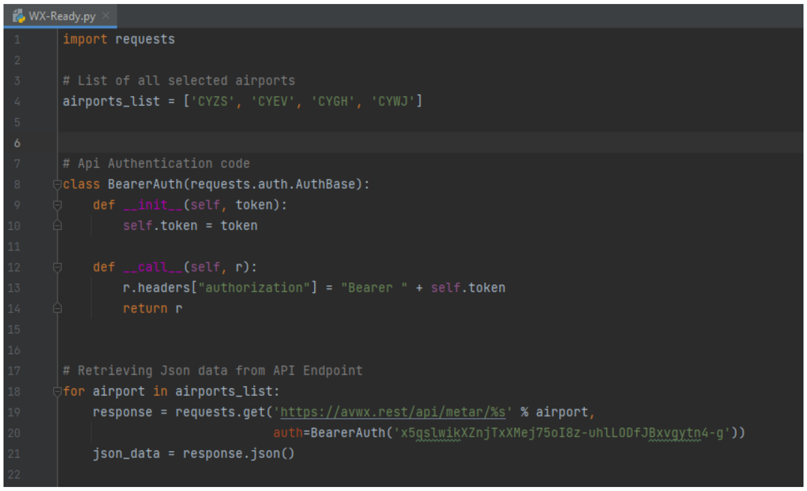
8.1 Backend METAR Info Retrieval “GET request.”

The backend component is the code that runs to retrieve the weather observations of the specified airports. This backend code uses API calls to an aviation Weather REST API service. The METAR has its specific endpoint shown in figure 9 and figure 10.



https://avwx.rest/api/metar/cyzs

FIGURE 9: METAR API Endpoint



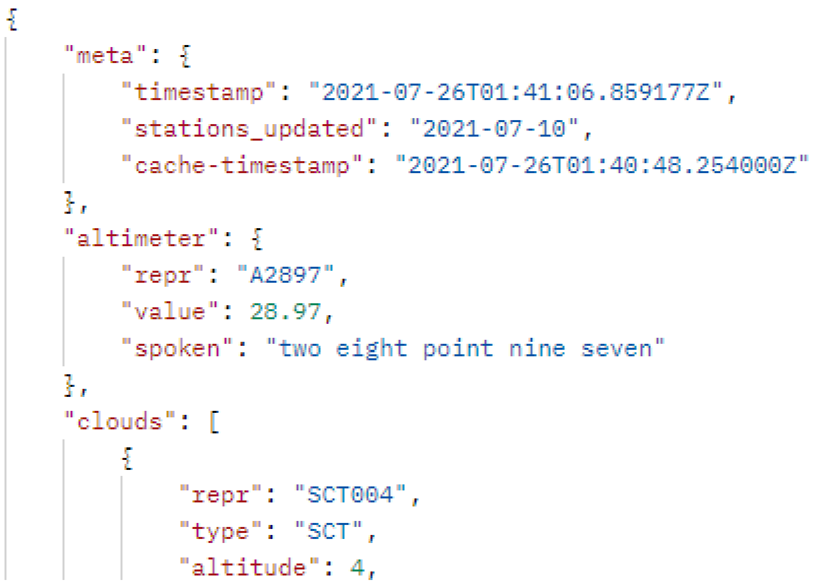
```python
import requests

# List of all selected airports
airports_list = ['CYZS', 'CYEV', 'CYGH', 'CYWJ']


# Api Authentication code
class BearerAuth(requests.auth.AuthBase):
    def __init__(self, token):
        self.token = token

    def __call__(self, r):
        r.headers["authorization"] = "Bearer " + self.token
        return r


# Retrieving Json data from API Endpoint
for airport in airports_list:
    response = requests.get('https://avwx.rest/api/metar/%s' % airport,
                            auth=BearerAuth('x5qslwikXZnjTxXMej75oI8z-uhlLODfJBxvgytn4-g'))
    json_data = response.json()
```

FIGURE 10: Python Code used to request weather data in.

The airport list contains all the airports the crew inputs into the UI. We then authenticate to the endpoint using the bearer token method. After the authentication, we modify the server response to a JSON format.

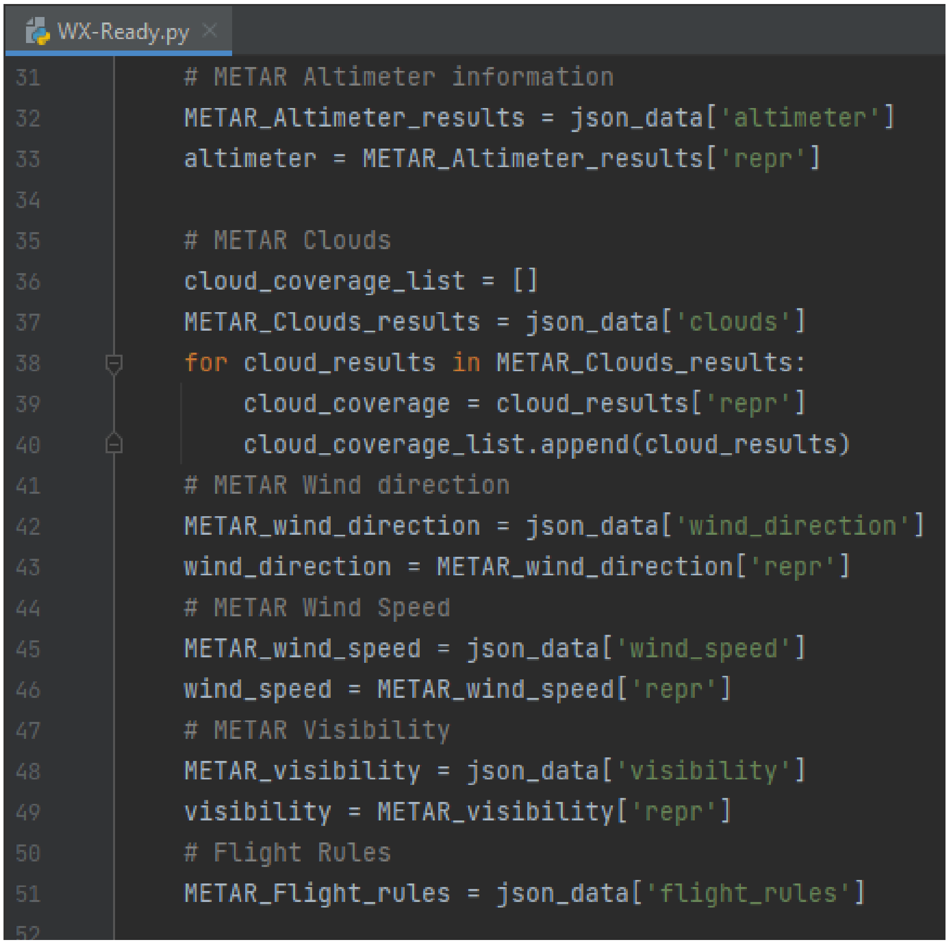
8.2 Backend METAR Info Retrieval “JSON data consumption.”

Now that our response is in JSON format as shown in figure 11, we place each weather observation in its variable, as shown in figure 12.



```
{
    "meta": {
        "timestamp": "2021-07-26T01:41:06.859177Z",
        "stations_updated": "2021-07-10",
        "cache-timestamp": "2021-07-26T01:40:48.254000Z"
    },
    "altimeter": {
        "repr": "A2897",
        "value": 28.97,
        "spoken": "two eight point nine seven"
    },
    "clouds": [
        {
            "repr": "SCT004",
            "type": "SCT",
            "altitude": 4,
```

FIGURE 11: METAR weather data in JSON format.

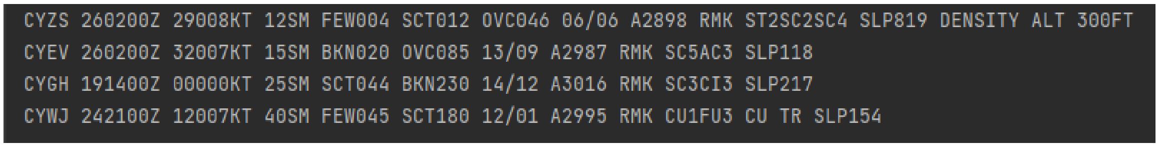


```
# METAR Altimeter information
METAR_Altimeter_results = json_data['altimeter']
altimeter = METAR_Altimeter_results['repr']

# METAR Clouds
cloud_coverage_list = []
METAR_Clouds_results = json_data['clouds']
for cloud_results in METAR_Clouds_results:
    cloud_coverage = cloud_results['repr']
    cloud_coverage_list.append(cloud_results)
# METAR Wind direction
METAR_wind_direction = json_data['wind_direction']
wind_direction = METAR_wind_direction['repr']
# METAR Wind Speed
METAR_wind_speed = json_data['wind_speed']
wind_speed = METAR_wind_speed['repr']
# METAR Visibility
METAR_visibility = json_data['visibility']
visibility = METAR_visibility['repr']
# Flight Rules
METAR_Flight_rules = json_data['flight_rules']
```

FIGURE 12: Python Code to dissect JSON data into variables.

8.3 Backend Weather Info Display In METAR Format

We can display the information gathered to the crew by printing the variables while respecting the International Civil Aviation Organization ICAO METAR format clearly shown in figure 13.

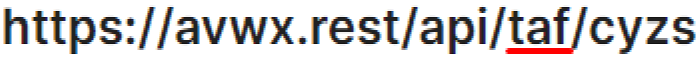


```
CYZS 260200Z 29008KT 12SM FEW004 SCT012 OVC046 06/06 A2898 RMK ST2SC2SC4 SLP819 DENSITY ALT 300FT
CYEV 260200Z 32007KT 15SM BKN020 OVC085 13/09 A2987 RMK SC5AC3 SLP118
CYGH 191400Z 00000KT 25SM SCT044 BKN230 14/12 A3016 RMK SC3CI3 SLP217
CYWJ 242100Z 12007KT 40SM FEW045 SCT180 12/01 A2995 RMK CU1FU3 CU TR SLP154
```

FIGURE 13: METAR weather info displayed in ICAO format.

8.4 Backend TAF Info Retrieval “GET request.”

This section of the backend code is the same as the METAR section. The only difference is that the TAF section uses its endpoint, as shown in figure 14.

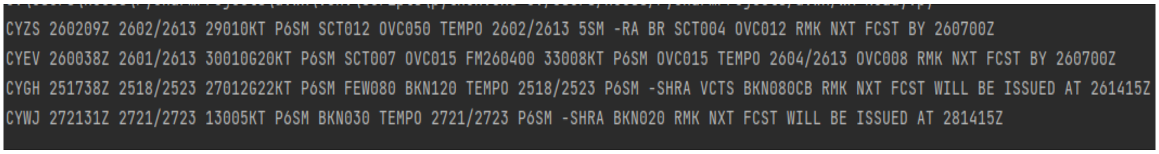


https://avwx.rest/api/taf/cyzs

FIGURE 14: TAF API Endpoint.

8.5 Backend Weather Info Display In TAF Format

In figure 15 We display the TAF information retrieved to the crew by printing the variables while respecting the ICAO TAF format.

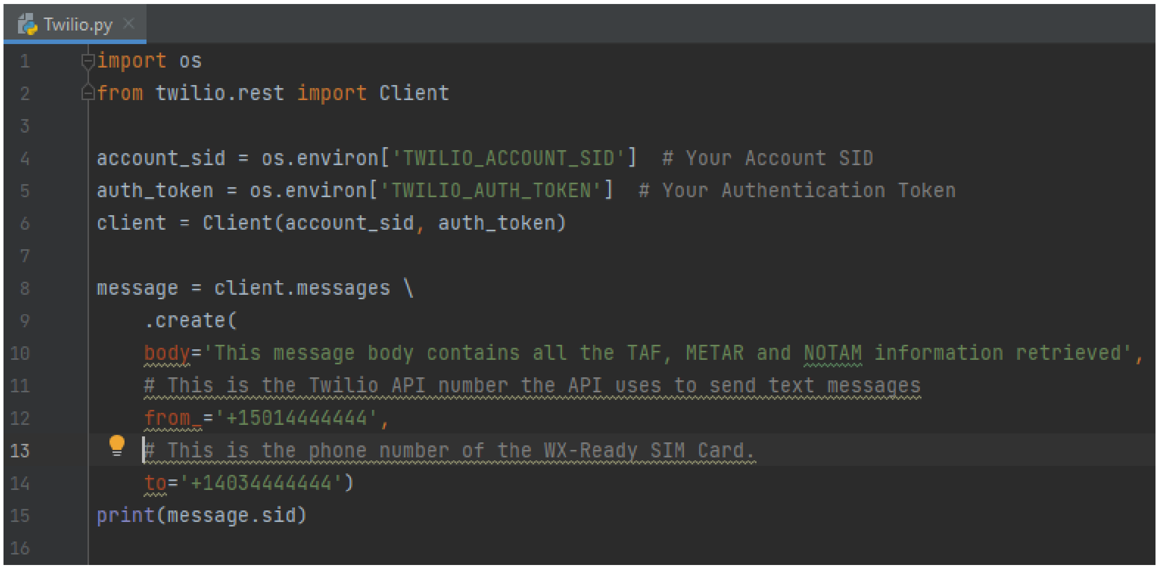


```
CYZS 260209Z 2602/2613 29010KT P6SM SCT012 OVC050 TEMPO 2602/2613 5SM -RA BR SCT004 OVC012 RMK NXT FCST BY 260700Z
CYEV 260038Z 2601/2613 30010G20KT P6SM SCT007 OVC015 FM260400 33008KT P6SM OVC015 TEMPO 2604/2613 OVC008 RMK NXT FCST BY 260700Z
CYGH 251738Z 2518/2523 27012G22KT P6SM FEW080 BKN120 TEMPO 2518/2523 P6SM -SHRA VCTS BKN080CB RMK NXT FCST WILL BE ISSUED AT 261415Z
CYWJ 272131Z 2721/2723 13005KT P6SM BKN030 TEMPO 2721/2723 P6SM -SHRA BKN020 RMK NXT FCST WILL BE ISSUED AT 281415Z
```

FIGURE 15: TAF weather info displayed in ICAO format.

8.6 Backend Sending Weather Data through Cellular Network

Once we have gathered all the weather data, the Twilio messaging service will reply to the same phone number that requested the info shown in figure 16.

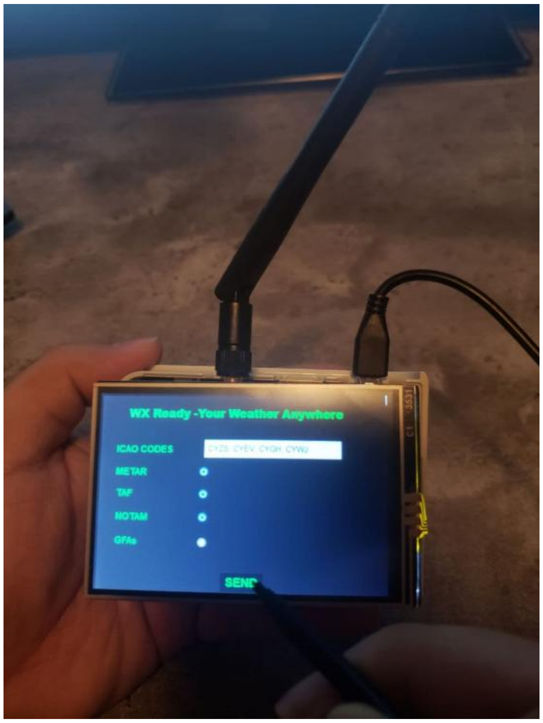


```python
import os
from twilio.rest import Client

account_sid = os.environ['TWILIO_ACCOUNT_SID']  # Your Account SID
auth_token = os.environ['TWILIO_AUTH_TOKEN']  # Your Authentication Token
client = Client(account_sid, auth_token)

message = client.messages \
    .create(
    body='This message body contains all the TAF, METAR and NOTAM information retrieved',
    # This is the Twilio API number the API uses to send text messages
    from_='+15014444444',
    # This is the phone number of the WX-Ready SIM Card.
    to='+14034444444')
print(message.sid)
```

FIGURE 16: Weather data transmission via SMS using Twilio API.

9. THE WORKING PROTOTYPE

This section shows how WX-Ready is operating in figure 17, figure 18, figure 19 and figure 20.

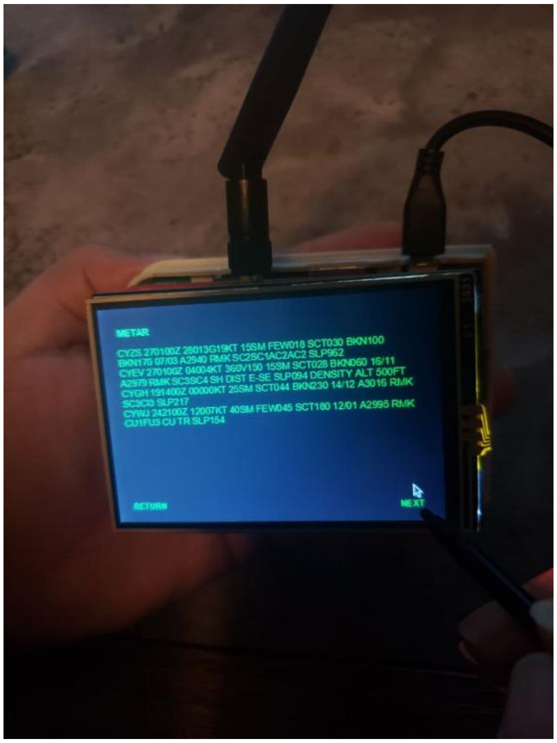


WX Ready -Your Weather Anywhere

ICAO CODES: CYZS CYEV CYGH CYWJ

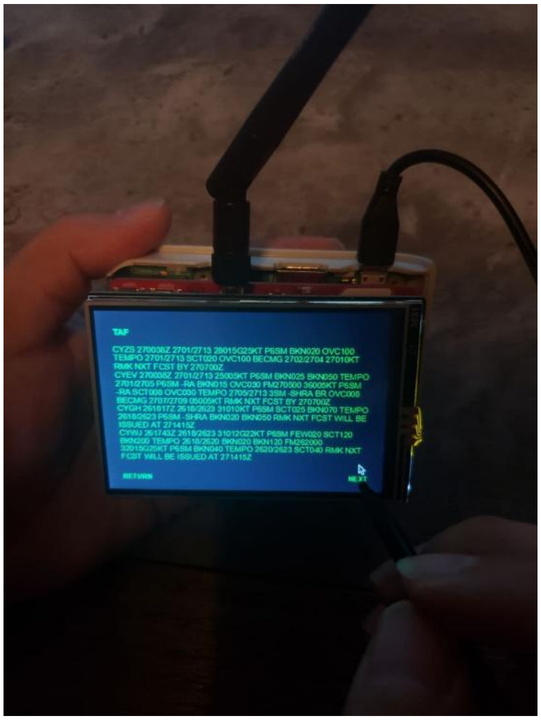
METAR ○
TAF ○
NOTAM ○
GFAs ○

FIGURE 17: WX-Ready main screen.



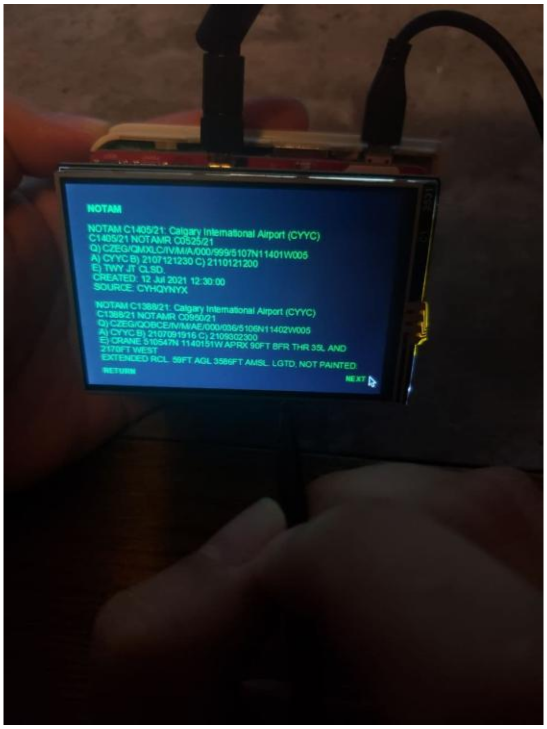
```
METAR
CYZS 270100Z 28013G19KT 15SM FEW018 SCT030 BKN100
BKN170 07/03 A2940 RMK SC2SC1AC2AC2 SLP962
CYEV 270100Z 04004KT 360V150 15SM SCT028 BKN060 16/11
A2979 RMK SC3SC4 SH DIST E-SE SLP094 DENSITY ALT 500FT
CYGH 191400Z 00000KT 25SM SCT044 BKN230 14/12 A3016 RMK
SC3C8 SLP217
CYWJ 242100Z 12007KT 40SM FEW045 SCT180 12/01 A2995 RMK
CU1FU3 CU TR SLP154
```

FIGURE 18: WX-Ready METAR screen.



```
CYZS 270038Z 2701/2713 28015G25KT P6SM BKN020 OVC100
TEMPO 2701/2713 SCT020 OVC100 BECMG 2702/2704 27010KT
RMK NXT FCST BY 270700Z
CYEV 270038Z 2701/2713 25005KT P6SM BKN025 BKN050 TEMPO
2701/2705 P6SM -RA BKN015 OVC030 FM270500 36005KT P6SM
-RA SCT008 OVC030 TEMPO 2705/2713 3SM -SHRA BR OVC005
BECMG 2707/2709 05005KT RMK NXT FCST BY 270700Z
CYGH 261817Z 2618/2623 31010KT P6SM SCT025 BKN070 TEMPO
2618/2623 P6SM -SHRA BKN020 BKN050 RMK NXT FCST WILL BE
ISSUED AT 271415Z
CYWJ 261743Z 2618/2623 31012G22KT P6SM FEW020 SCT120
BKN200 TEMPO 2618/2620 BKN020 BKN120 FM262000
32015G25KT P6SM BKN040 TEMPO 2620/2623 SCT040 RMK NXT
FCST WILL BE ISSUED AT 271415Z
```

FIGURE 19: WX-Ready TAF screen.



```
NOTAM C1405/21: Calgary International Airport (CYYC)
C1405/21 NOTAMR C0525/21
Q) CZEG/QMXLC/IV/M/A/000/999/5107N11401W005
A) CYYC B) 2107121230 C) 2110121200
E) TWY JT CLSD.
CREATED: 12 Jul 2021 12:30:00
SOURCE: CYHQYNYX

NOTAM C1388/21: Calgary International Airport (CYYC)
C1388/21 NOTAMR C0950/21
Q) CZEG/QOBCE/IV/M/AE/000/036/5106N11402W005
A) CYYC B) 2107091916 C) 2109302300
E) CRANE 510547N 1140151W APRX 90FT BFR THR 35L AND
2170FT WEST
EXTENDED RCL. 59FT AGL 3586FT AMSL. LGTD, NOT PAINTED.
```

FIGURE 20: WX-Ready NOTAM screen.

10. FUTURE ENHANCEMENTS

WX-Ready is eventually intended to operate in the most remote areas in the world. To reach this goal, the apparatus needs to send the requests via the GSM network and through the Iridium satellite network when the GSM network is not available. Since it is based on the Raspberry PI hardware platform, the WX-Ready Apparatus can easily be connected to a Rockblock Mk2 Iridium modem shown in figure 21. This modem can send SMS messages through the Iridium Satellite Constellation via the SBD Short Burst Data communication protocol. Adding this capability to WX-Ready does not change the systems' architecture discussed above because all we do is sending the SMS message through a different network. The backend servers and processes remain the same.

FIGURE 21: RockBlock Satellite Modem.

11. CONCLUSION

The developed prototype performed its job correctly. In the future, we should add additional features to the software and upgrades to the hardware. Currently, the system can operate anywhere in Canada, provided it's within Cellular coverage. To have WX-Ready work virtually anywhere globally, we can add an Iridium transceiver that can communicate with the Iridium satellite constellation using the Short Burst Data protocol (SBD). Currently, we are getting the power from a USB connection; however, we can also add a battery to provide power to the device. WX-Ready costs less than 40$ and has an operating cost of 10$ per month. This apparatus is the cheapest way to retrieve aviation weather information in remote areas versus commercially available ACARS solutions with prices ranging from $200.000 to $400.000.

WX-Ready benefits airline operators in the Arctic and remote regions like Africa, South America, and Asia.

12. REFERENCES

Anaman, Kwabena & Quaye, Ruth & Owusu-Brown, Bernice. (2017). Benefits of Aviation Weather Services: A Review of International Literature. *Research in World Economy.* 8. 45-58. 10.5430/rwe.v8n1p45.

Aviation Weather Services (2009). Advisory Circular 00-45F, Change 2. *Federal Aviation Administration.*

Government of Canada, O. of the A. G. of C.(2017, May 16). *Report 6—Civil Aviation Infrastructure in the North—Transport Canada*. Www.oag-Bvg.gc.ca. https://www.oag-bvg.gc.ca/internet/English/parl_oag_201705_06_e_42228.html

International Civil Aviation Organization (ICAO).(2010, July).*Annex 3 to the Convention on International Civil Aviation: Meteorological Service for International Air Navigation, 20th ed.*

Schwartz, B., & Benjamin, S. G. (1995). A Comparison of Temperature and Wind Measurements from ACARS-Equipped Aircraft and Rawinsondes. *Weather and Forecasting, 10*(3), 528–544. https://doi.org/10.1175/1520-0434

Smith, M., Strohmeier, M., Lenders, V., & Martinovic, I. (2016). On the security and privacy of ACARS. *Integrated Communications Navigation and Surveillance (ICNS).* Published. https://doi.org/10.1109/icnsurv.2016.7486395

Susanto, A., & Meiryani.(2019).*System Development Method with The Prototype Method.* International Journal of Scientific and Technology Research8(07) pp. 142–143

Tamalet, S., Gobbo, G., Durand, F., & Deville, J. G. (2007). Acars router for remote avionics applications (US8484384B2).*United States Patent.* https://patents.google.com/patent/US8484384B2/en

INSTRUCTIONS TO CONTRIBUTORS

The *International Journal of Computer Science and Security (IJCSS)* is a refereed online journal which is a forum for publication of current research in computer science and computer security technologies. It considers any material dealing primarily with the technological aspects of computer science and computer security. The journal is targeted to be read by academics, scholars, advanced students, practitioners, and those seeking an update on current experience and future prospects in relation to all aspects computer science in general but specific to computer security themes. Subjects covered include: access control, computer security, cryptography, communications and data security, databases, electronic commerce, multimedia, bioinformatics, signal processing and image processing etc.

To build its International reputation, we are disseminating the publication information through Google Scholar, J-Gate, Docstoc, Scribd, Slideshare, Bibsonomy and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJCSS.

The initial efforts helped to shape the editorial policy and to sharpen the focus of the journal. Started with Volume 15, 2021, IJCSS is appearing with more focused issues. Besides normal publications, IJCSS intend to organized special issues on more focused topics. Each special issue will have a designated editor (editors) – either member of the editorial board or another recognized specialist in the respective field.

We are open to contributions, proposals for any topic as well as for editors and reviewers. We understand that it is through the effort of volunteers that CSC Journals continues to grow and flourish.

IJCSS LIST OF TOPICS
The realm of International Journal of Computer Science and Security (IJCSS) extends, but not limited, to the following:

- Authentication and authorization models
- Computer Engineering
- Computer Networks
- Cryptography
- Databases
- Image processing
- Operating systems
- Programming languages
- Signal processing
- Theory
- Communications and data security
- Bioinformatics
- Computer graphics
- Computer security
- Data mining
- Electronic commerce
- Object Orientation
- Parallel and distributed processing
- Robotics
- Software engineering

CALL FOR PAPERS

Volume: 15 - Issue: 5

i. Submission Deadline : August 31, 2021 **ii. Author Notification:** September 30, 2021

iii. Issue Publication: October 2021

CONTACT INFORMATION

Computer Science Journals Sdn BhD

B-5-8 Plaza Mont Kiara, Mont Kiara

50480, Kuala Lumpur, MALAYSIA

Phone: 006 03 6204 5627
Fax: 006 03 6204 5628
Email: cscpress@cscjournals.org

CSC PUBLISHERS © 2021
COMPUTER SCIENCE JOURNALS SDN BHD
B-5-8 PLAZA MONT KIARA
MONT KIARA
50480, KUALA LUMPUR
MALAYSIA

PHONE: 006 03 6204 5627
FAX: 006 03 6204 5628
EMAIL: cscpress@cscjournals.org