Science World Journal Vol. 20(No 3) 2025 https://dx.doi.org/10.4314/swj.v20i3.55
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

A COMPILER APPROACH TO PHARMACOVIGILANCE

*Mani Kitgwim Christopher, Davou Choji Nyap, Nachamada Vachaku Blamah, Nwosu Nkechi Peace, Zingdul Kenneth Ponnan

Department of Computer Science, Faculty of Computing, University of Jos, Nigeria

*Corresponding Author Email Address: kitgwimmani@gmail.com

ABSTRACT
Special-purpose compilers had been developed for specific problem domains, which mainly focused on improving certain aspects of the compiler. For instance, the NNVM compiler and the TensorFlow AXL compiler proffers solutions to improve the performance of machine learning algorithms by implementing parallel computing and code optimization to reduce time complexity. An obvious burden in data science/machine learning undertakings is the amount of time needed for data preprocessing. This study has a new look at compiler design using pharmacovigilance as a case study. The research developed a special-purpose compiler to be used in improving data preprocessing as applied to pharmacovigilance. Lexical analysis was applied to data preprocessing and hashing techniques in surveillance and case reporting. The dataset used in the study contains some demographic information of the patients, drugs prescribed, and reported adverse effects. The compiler was built using the Python programming language, and a random forest model was developed using 70% of the data as a training set while the remaining 30% was reserved for testing. The initial model performance in terms of accuracy in reporting adverse events was 0.08; however, after applying hashing techniques and adding the hash as an additional attribute to the dataset, a 1.0 (100%) accuracy was achieved.

Keywords: Compiler design; Pharmacovigilance; Adverse Drug Events (ADE); Data preprocessing; Lexical analysis; Machine learning; Hashing techniques

INTRODUCTION

In recent times, there has been adoption and application of compiler techniques in deep learning systems to achieve better performance, for instance, TensorFlow Accelerated Linear Algebra (XLA) compiler (Chris and Todd 2017; He, 2023) and the Neural Network Virtual Machine (NNVM) compiler (Richard et al., 2017; Tiwari et al., 2024). With a growing interest in the field of data science and Artificial Intelligence (AI), the relevance of Big Data cannot be overemphasized. However, despite the increased attention given to this aspect of technology, the problem of data quality and reliability persists. It is well known that 80 - 90% of the time spent on machine learning development is data preparation (Stonebraker and Rezig, 2019). This issue is common because of the heterogeneity of data sources, resulting in data variety (a major challenge of big data). As a result, compilers are being modified to increase performance, especially in the aspect of data preprocessing.

Expert systems built on machine learning techniques rely on data quality to achieve their goal (Eunsuk et al., 2017; Straub, 2021). The challenges associated with data quality for the development of expert systems are encountered mostly in classification tasks in which the distribution of classes or labels in a given dataset is not uniform, and the common approach used in solving this problem is by oversampling or under sampling (Roweida et al., 2020). One of the focal points of this study is to minimize the need for data preprocessing by improving data quality using a compiler approach.

The use of data mining and machine learning for drug safety surveillance in the field of clinical pharmacology is gaining more attention (Bete et al., 2018). However, just like other expert systems that rely on a variety of data, there is a problem of data quality and a high need for data preprocessing (Stonebraker and Rezig, 2019). This data is usually collected from doctors and pharmacists in a health care facility. The main drawback of this data is that its correctness has to be checked, as there may be some interrelated issues such as duplications and so on (Budach et al., 2022; Harpaz, 2016).

There is a clear attention to the usefulness of Artificial Intelligence and Machine learning in pharmacovigilance (PV) (Murali et al., 2019; Hussain et al., 2021; Sessa, 2021). The kind of data needed to be processed for PV is diverse and varies greatly in terms of quality and quantity. Data from the Individual Case Study Report (ICSR) should be consistent with the data from the clinical trial, but this is not true in reality (Sessa, 2021). Obtaining the right report from patients, caregivers, or Health Care Professionals is inherently complicated. Medical Literature Monitoring (MLM), which is peer-reviewed, is another important source of safety data, and the peer-review method should assure the data's credibility. However, the information supplied is typically limited to what is found in the paper.

Social media provides a huge amount of PV data, which cannot be verified easily. It has been proven that data, which can be relevant for PV purposes, are more reliable when they emanate from private media platforms dedicated to Health Practitioners (for instance, online professional communities) than those obtained from open platforms like Facebook and Twitter, where everyone can make posts (Sessa, 2021). This clearly shows that having a divergent data source for pharmacovigilance purposes presents integrity concerns. In fact, some of the data gotten from social media is either false or even willfully misleading.

Wearable Devices (Fitbits, Garmins, iPhones, etc.) that can generate millions or billions of data points about heart rate, sleep/wake patterns, activity, and, in some cases, oxygen saturation and blood sugar can provide a reliable source of data. The Apple watch, released in 2018, Electrocardiography ECG App is an example of a wearable device that is a more reliable data source for PV. Needless to say, this approach has sparked a lot of debate and the writing of several papers on the pros and cons of using such devices (L'Hommedieu et al., 2019; Perez et al., 2019; Whelan et al., 2019).

The fundamental issue in machine-aided PV case processing, as in many other fields where natural language is used, is that the data to be processed might be both structured and unstructured, requiring separate processing and analysis (Marco et al.,2021). A tremendous amount of time and effort is being spent on data

Science World Journal Vol. 20(No 3) 2025 https://dx.doi.org/10.4314/swj.v20i3.55
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

preprocessing (Stonebraker and Rezig, 2019). Well-labeled data can be obtained from relational databases and thus provides little or no preprocessing challenge. However, data captured as narratives are unstructured and pose a greater challenge during preprocessing because the relevant information may be hidden deeply within the message or may not exist; in both cases, there would be a dissipation of energy for preprocessing. Sometimes unstructured data is even more complicated because of wrong spellings and abbreviations, for instance, a typical Facebook post or a Twitter post. The more unstructured the data, the more "intelligence" is required to process it in a meaningful way (Wong, 2018).

Another important concern is data veracity. Having a variety of PV data from different sources poses a challenge of acceptability and of the authenticity of the data source. Even where there is success in preprocessing, meaningful inferences can only be drawn from a legitimate data source (entries from professional or first-hand from caregiver or patient experiencing Adverse Drug Events, ADE).

MATERIALS AND METHODS

The system contains two major aspects which are the lexical and syntax analysis (compiler approach) aspect for tokenization of prescription records and the machine learning (ML) aspect, which performs the pharmacovigilance and renders real-time feedback to the dashboard. The concept of the system is similar to outlier detection but in this case, the outliers are unprecedented adverse drug reactions that require further investigation and, in some cases, immediate intervention. In the compiler design stage, a syntax tree containing grammatical rules would be created, which defines acceptable expressions with their respective inferences. Regular expression (ReGex) would be used to define acceptable tokens. The preceding step is the creation and deployment of the surveillance reporting system, which would be the primary input source of adverse drug reactions in the system. Every entry would be organized into a single code and hashed. The code and the hashes would be routed into the pharmacovigilance repository, where the intelligent agent would classify such an entry for analysis. The idea of classification is to identify anomalies in the surveillance report and flag off for the system. The compiler would be used to tokenize input from the surveillance system to provide high-precision data from which error tokens can be identified, which constitute interesting cases, and thus, drug batch numbers, when identified, can be reported for immediate action. The compiler approach is considered in this study to be superior to NLP because of the concept of optimization, which is fundamentally performance-oriented in terms of processing needs and also as a result of tokenization, which directly aligns with breaking down components of a typical prescription (drug, dosage, frequency, and route of administration) (Kovac et al., 2022).

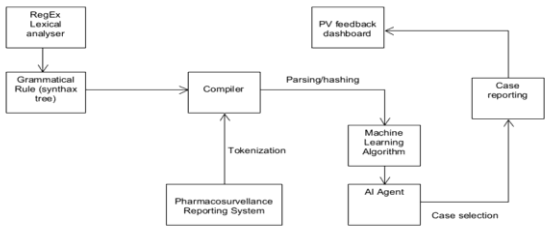

Figure 1: The system architecture (showing compiler phases and machine learning phases)

Figure 1 shows the design and implementation flow for all the stages. The lexical analyzer defines the acceptable tokens in the system, which are standard drug abbreviations, dosage forms, and adverse symptoms. These form the basis for developing grammatical rules, a process for the development of a syntax tree. These form the building block for the compiler. The pharmacovigilance reporting system contains input forms where reporting actors input the details obtained from the patient. These details are converted into a more precise format in a process known as tokenization. Each entry is termed a 'capsule', a special term used in this study to describe a tuple of a surveillance unit, including its respective hash and other medical, demographic, and supporting details provided during PV reporting. A capsule is the basic unit (row) that forms the input for the machine learning. The ML uses clustering and classification to create the ML agent, which is responsible for isolating interesting cases that form the focal input for pharmacovigilance (examples of interesting cases are new adverse effects or unexpected drug reactions, including drug-drug interactions). After successful case selection, the Capsules selected are used to create more informed feedback, which can be used by the professionals for immediate action where needed. This can reveal unsuspecting hidden patterns for the pharmacovigilance reporting system, which is not easily obtainable in a manual system.

Description of the dataset and the data source

In order to achieve the implementation as described in the methodology of this study, the datasets were downloaded from the preceding links provided and explored in the Jupiter notebook environment. Below are several outputs generated from the data set. One of the important aspects of this data set is the various attributes that are relevant to the study, for instance, drug name, medical condition, and normal symptoms. Normal symptom is expected symptoms based on the pharmacology of the drugs, which form the basis for the surveillance report in which the compiler identifies any symptom that is not listed as part of the normal to be flagged as a potential unprecedented adverse effect, which is needed for further investigation. From the unprecedented symptoms, the compiler will produce tokens that are representatives of the drug and the side effects; these tokens will provide independent entities that will be further analyzed to reveal insights.

Data source = https://www.kaggle.com/datasets/jithinanievarghese/drugs-side-effects-and-medical-condition?resource=download

	drug_name	medical_condition	side_effects	generic_name	drug_classes	brand_names	activity	rx_otc	pregnancy_category	csa	...	Un
0	doxycycline	Acne	(hives, difficult breathing, swelling in your ...	doxycycline	Miscellaneous antimalarials, Tetracyclines	Acticlate, Adoxa CK, Adoxa Pak, Adoxa TT, Alod...	87%	Rx	D	N	...	
1	spironolactone	Acne	hives ; difficulty breathing; swelling of your...	spironolactone	Aldosterone receptor antagonists, Potassium-sp...	Aldactone, CaroSpir	82%	Rx	C	N	...	
2	minocycline	Acne	skin rash, fever, swollen glands, flu-like sym...	minocycline	Tetracyclines	Dynacin, Minocin, Minolira, Solodyn, Ximino, V...	48%	Rx	D	N	...	
3	Accutane	Acne	problems with your vision or hearing;	isotretinoin (oral)	Miscellaneous antineoplastics, Miscellaneous	NaN	41%	Rx	X	N	...	

Figure 2: First five records from the dataset

RESULTS

Table I: Evaluation Scores for both control and test

	Accuracy Score	Precision	Recall	F1-score	AUC
Control set	**0.08416666**	**0.0842**	**0.1325**	**0.1036**	**0.2174**
Test set	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table I shows that the hash attribute contributed immensely in improving the performance of the decision tree model. To further explain the outcome, a confusion matrix in Figure 2 was used to present the outcome of the test set with detailed performance matrices. Specifically, presenting perfect accuracy, precision, recall, f1-score, and AUC as presented in the table. The dramatic jump in performance can be explained by the normalization of the dataset, with hashing that gives a very precise reference that encapsulates all the nuances in the raw data.

With reference to the objectives of the study, the entire spectrum proposed by the system was performed. The compiler process ensured that input files were read using the lexical structure defined, and the parsing procedure ensured that accurate grammatical rules were followed. The compiler was instrumental in the data preprocessing process to ensure that attributes of the input set were verified and validated before invoking the hashing algorithm. The output, a CSV file which contains medical diagnosis, age, gender, and reported drug side effects, was primary, and the hash of values was appended as additional variables. Two sets of the dataset were defined as control (containing age and gender as a training subset) and test, which contains the hash attribute in addition. A decision tree model was used in both cases, with drug side effects as the outcome variable. Accuracy scores of 0.084 and 1.00 were the result of the control and the test sets, respectively (Fig. 2), showing that the compiler-based approach and hashing technique provide optimal performance for machine learning as well as better outcomes in pharmacovigilance surveillance systems.

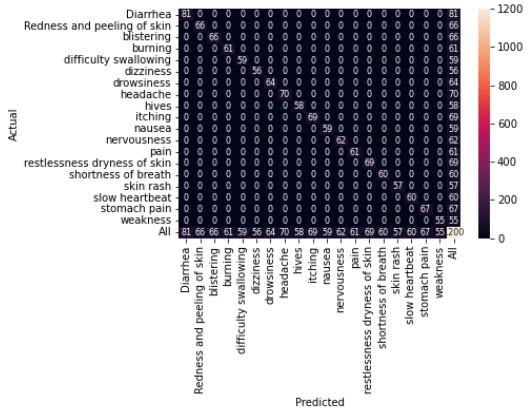

Figure 1: Confusion Metrix

DISCUSSION

The current dispensation of information and communication technology, powered by the field of computer science, invokes a trend of disruptive innovations. This spans through all levels of computational sciences and paves the way for adopting fundamental principles of computer science, one of which is the old, good compiler construction. The advancing field of data science has applied compiler techniques in improving processes and procedures, notably, the TensorFlow XLA compiler (Leary, 2017) and the NNVM compiler (Richard, 2017).

These solutions focused on performance optimization in terms of processing time and storage. This study considers the applicability of compiler techniques in the area of data preparation, which provides an opportunity for innovation because, according to Stonebraker (2019), data processing consumes over 80% of the time required for data science and machine learning projects. To put this study into a better perspective, pharmacovigilance was used as a case study. This case study was suitable because pharmacovigilance datasets are varied and intricate details could lodge within a highly unstructured dataset (Marco,2021).

A compiler was developed using Python programming language to perform lexical analysis of text files and other flat files in order to develop a more comprehensive CSV file to be used for machine learning and pharmacovigilance purposes. The compiler also invoked a hashing function to create additional attributes that easily map specific attributes of a precise reported symptom. The outcome of the study shows that the compiler approach aided data preprocessing and also yielded higher accuracy when tested using a decision tree model.

The study is limited by time and resources to deploy the solution for real-time testing. However, within the scope of the study, the outcome of the study is a step-in compiler-based disruptive innovations. The compiler technique is fundamental to computer science and should be considered alongside data structures, algorithm optimization and memory management in revolutionizing computational sciences amidst the ever-dynamic disruptions today. Due to the fact that data preprocessing consumes a huge fraction of the machine learning cycle, there should be automation of the process. This study shows that a compiler approach to data preprocessing can improve the accuracy score of models in addition to a more precise visualization during data exploration. More investment of time and technical resources can aid in achieving an optimal industry-specific computational solution that can advance humanity as a whole.

REFERENCES

Asha Kiranmai, S., & Jaya Laxmi, A. (2018). Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. *Protection and Control of Modern Power Systems, 3*(1), 1–12. https://pcmp.springeropen.com/articles/10.1186/s41601-018-0103-3

Balian, J. D., Wherry, J. C., Malhotra, R., & Perentesis, V. (2010). Roadmap to risk evaluation and mitigation strategies. *Therapeutic Advances in Drug Safety, 1*(1), 21–38. https://doi.org/10.1177/204209861038141

Bate, A., Reynolds, R. F., & Caubel, P. (2018). The hope, hype and reality of Big Data for pharmacovigilance. *Therapeutic Advances in Drug Safety, 9*, 5–11. https://doi.org/10.1177/2042098617736422

Bisong, E. (2019). Introduction to Scikit-learn. In *Building machine learning and deep learning models on Google cloud platform* (pp. 215–229). Apress.

Boland, M. R., & Tatonetti, N. P. (2015). Are all vaccines created equal? Using electronic health records to discover

vaccines associated with clinician-coded adverse events. *AMIA Summits on Translational Science Proceedings, 2015*, 196.

Chandel, S., Jain, N., Joshi, A., Sonawane, R., Sharma, A., & Chandel, S. (2014). Signal detection – an imperative activity of pharmacovigilance. *International Journal of Pharmaceutical Sciences Review and Research, 28*, 95–100.

Leary, C., & Wang, T. (2017, February). XLA: TensorFlow, compiled! *TensorFlow Dev Summit*.

Lee, C. Y., & Chen, Y.-P. P. (2019). Machine learning on adverse drug reactions for pharmacovigilance. *Drug Discovery Today, 24*(7), 1332–1343. https://doi.org/10.1016/j.drudis.2019.03.003

Duggirala, H. J., Tonning, J. M., & Smith, E. (2016). Use of data mining at the Food and Drug Administration. *Journal of the American Medical Informatics Association, 23*(2), 428–434. https://doi.org/10.1093/jamia/ocv063

Choi, E., Heo, C., & Park, F. C. (2017). Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. *Expert Systems with Applications, 83*, 187–205. https://doi.org/10.1016/j.eswa.2017.04.030

Ghosh, R. E., Padmanabhan, S., & Williams, R. (2018). Including primary care data from multiple software systems in a data linkage programme: Results from expanding the Clinical Practice Research Datalink (CPRD). *Pharmacoepidemiology and Drug Safety, 27*(S2), 89. https://doi.org/10.1002/pds.4570

Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., & Greene, C. S. (2016). Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics, 17*(1), 33–42. https://doi.org/10.1093/bib/bbv087

Goyvaerts, J. (2017). *Regular expression tutorial: Learn how to use regular expressions*. Regular-Expressions.info. http://www.regular-expressions.info

Harpaz, R. (2014). Text mining for adverse drug events: The promise, challenges, and state of the art. *Drug Safety, 37*(10), 777–790. https://doi.org/10.1007/s40264-014-0218-z

Hussain, R. (2021). Big data, medicines safety and pharmacovigilance. *Journal of Pharmaceutical Policy and Practice, 14*(48). https://doi.org/10.1186/s40545-021-00329-4

Létinier, L., et al. (2021). Artificial intelligence for unstructured healthcare data: Application to coding of patient reporting of adverse drug reactions. *Clinical Pharmacology & Therapeutics*. https://doi.org/10.1002/cpt.2266

L'Hommedieu, M., L'Hommedieu, J., Begay, C., Schenone, A., Dimitropoulou, L., Margolin, G., Falk, T., Ferrara, E., Lerman, K., & Narayanan, S. (2019). Lessons learned: Recommendations for implementing a longitudinal study using wearable and environmental sensors in a health care organization. *JMIR mHealth and uHealth, 7*(12), e13305. https://preprints.jmir.org/preprint/13305

Marco, A., Xavier, F., Zurab, K., Flemming, K. J., Nicolas, T., Miranda, D., Errietta, E., & Essam. (2021). Artificial intelligence (AI) in pharmacovigilance: Do we really need it? European CRO Federation (EUCROF).

Mokdad, A. H., Hotez, P. J., & Orenstein, W. A. (2021). We have to get it right: Ensuring success. *EClinicalMedicine, 31*, 100690. https://doi.org/10.1016/j.eclinm.2020.100690

Moore, T. J., & Furberg, C. D. (2015). Electronic health data for post-market surveillance: A vision not realized. *Drug Safety, 38*(7), 601–610. https://doi.org/10.1007/s40264-015-0305-9

Murali, S. (2019). Artificial intelligence in pharmacovigilance: Practical utility. *Indian Journal of Pharmacology, 51*(6), 373–376. https://journals.lww.com/iphr/toc/2019/51060

Okazaki, N. (2014). CRFsuite: A fast implementation of conditional random fields (CRFs). http://www.chokkan.org/software/crfsuite/

Pawloski, P., Cusick, D., & Amborn, L. (2012). Risk evaluation and mitigation strategies: A decade of experience. *American Journal of Health-System Pharmacy, 69*(1), 49–54. https://doi.org/10.1093/ajhp/zxaa177

Perez, M. V., Mahaffey, K. W., Hedlin, H., Rumsfeld, J. S., Garcia, A., Ferris, T., Balasubramanian, V., Russo, A. M., Rajmane, A., Cheung, L., Hung, G., Lee, J., Kowey, P., Talati, N., Nag, D., Gummidipundi, S. E., Beatty, A., Hills, M. T., Desai, S., ... Turakhia, M. P. (2019). Large-scale assessment of a smartwatch to identify atrial fibrillation. *New England Journal of Medicine, 381*(20), 1909–1917. DOI: 10.1056/NEJMoa190118 https://doi.org/10.1056/NEJMoa1901183

Price, J. (2016). What can big data offer the pharmacovigilance of orphan drugs? *Clinical Therapeutics, 38*(12), 2433–2445. https://doi.org/10.1016/j.clinthera.2016.11.009

Ghosh, R. E., Crellin, E., Bates, S., Dobson, K., Myles, P., & Williams, R. (2019). How Clinical Practice Research Datalink data are used to support pharmacovigilance. *Therapeutic Advances in Drug Safety, 10*, 1–7. https://doi.org/10.1177/2042098619854010

Richard, W., Lane, S., & Vikram, A. (2017). DLVM: A modern compiler infrastructure for deep learning systems. *arXiv*. https://doi.org/10.48550/arXiv.1711.03016

Ritter, A., Clark, S., & Etzioni, O. (2011). Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 1524–1534).

Roweida, M., Jumanah, R., & Malak, A. (2020). Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *2020 International Conference on Innovative Computing and Communication (ICICS)* (pp. 239556–239562). IEEE. https://doi.org/10.1109/ICICS49469.2020.239556

Sarker, A., & Gonzalez, G. (2014). Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*. Advance online publication. https://doi.org/10.1016/j.jbi.2014.11.002

Sessa, M. (2021). Artificial intelligence in pharmacoepidemiology: A systematic review. Part 2 – Comparison of the performance of artificial intelligence and traditional pharmacoepidemiological techniques. *Frontiers in Pharmacology, 11*, 568659. https://doi.org/10.3389/fphar.2020.568659

Stonebraker, M., & Rezig, E. K. (2019). Machine learning and big data: What is important? *IEEE Data Engineering Bulletin, 42*(1), 1–5.

Trifirò, G., Sultana, J., & Bate, A. (2018). From big data to smart data for pharmacovigilance: The role of healthcare databases and other emerging sources. *Drug Safety, 41*, 143–149. https://doi.org/10.1007/s40264-017-0592-4

Visacri, M. B., de Souza, C. M., Sato, C. M. S., et al. (2015). Adverse drug reactions and quality deviations monitored by spontaneous reports. *Saudi Pharmaceutical Journal, 23*, 130–137. https://doi.org/10.1016/j.jsps.2014.07.002

Weng, S. F., Reps, J., Kai, J., et al. (2017). Can machine learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE, 12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

Whelan, M. E., Orme, M. W., Kingsnorth, A. P., Sherar, L. B., Denton, F. L., & Esliger, D. W. (2019). Examining the use of glucose and physical activity self-monitoring technologies in individuals at moderate to high risk of developing type 2 diabetes: Randomized trial. *JMIR mHealth and uHealth,* 7(10), e14195. https://doi.org/10.2196/14195

Wong, A. (2018). Natural language processing and its implications for the future of medication safety: A narrative review of recent advances and challenges. *Pharmacotherapy, 38*(8), 822–841. https://doi.org/10.1002/phar.2151

Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., ... & Harmouch, H. (2022). The effects of data quality on machine learning performance. *arXiv*. https://doi.org/10.48550/arXiv.2207.14529

Kovac, M., Brcic, M., Krajna, A., & Krleza, D. (2022). Towards intelligent compiler optimization. In *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)* (pp. 948–953). IEEE. https://doi.org/10.23919/MIPRO55190.2022.9803621

He, X. (2023). Accelerated linear algebra compiler for computationally efficient numerical models: Success and potential area of improvement. *PLOS ONE, 18*(2), e0282265. https://doi.org/10.1371/journal.pone.0282265

Tiwari, A., Fava-Rivi, C., Bandar, S. M., & Makandar, S. (2024). Optimizing machine learning models using Tensor Virtual Machine for embedded CPUs. In *2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)* (pp. 1–6). IEEE. https://doi.org/10.1109/SCOPES60114.2024.10584792

Straub, J. (2021). Machine learning performance validation and training using a 'perfect' expert system. *MethodsX, 8*, 101477. https://doi.org/10.1016/j.mex.2021.101477