Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Intelligence-Based Medicine 3-4 (2020) 100009

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.journals.elsevier.com/intelligence-based-medicine

Agile clinical research: A data science approach to scrumban in clinical medicine

Howard Lei [a,b,*], Ryan O'Connell [c], Louis Ehwerhemuepha [a,b,d], Sharief Taraman [a,b,e], William Feaster [a,b], Anthony Chang [a,b]

[a] CHOC Children's Hospital, Orange, CA, USA
[b] The Sharon Disney Lund Medical Intelligence and Innovation Institute (MI3), USA
[c] University of California-Irvine, Department of Pathology, USA
[d] Chapman University School of Computational and Data Science, Orange, CA, USA
[e] Department of Pediatrics, University of California-Irvine, School of Medicine, USA

ARTICLE INFO

Keywords:
Agile
Scrumban
Minimal Viable Model
Cloud Computing
Predictive model
Amazon Web Services

ABSTRACT

The COVID-19 pandemic has required greater minute-to-minute urgency of patient treatment in Intensive Care Units (ICUs), rendering the use of Randomized Controlled Trials (RCTs) too slow to be effective for treatment discovery. There is a need for agility in clinical research, and the use of data science to develop predictive models for patient treatment is a potential solution. However, rapidly developing predictive models in healthcare is challenging given the complexity of healthcare problems and the lack of regular interaction between data scientists and physicians. Data scientists can spend significant time working in isolation to build predictive models that may not be useful in clinical environments.

We propose the use of an agile data science framework based on the Scrumban framework used in software development. Scrumban is an iterative framework, where in each iteration larger problems are broken down into simple do-able tasks for data scientists and physicians. The two sides collaborate closely in formulating clinical questions and developing and deploying predictive models into clinical settings. Physicians can provide feedback or new hypotheses given the performance of the model, and refinement of the model or clinical questions can take place in the next iteration.

The rapid development of predictive models can now be achieved with increasing numbers of publicly available healthcare datasets and easily accessible cloud-based data science tools. What is truly needed are data scientist and physician partnerships ensuring close collaboration between the two sides in using these tools to develop clinically useful predictive models to meet the demands of the COVID-19 healthcare landscape.

1. Limitations of traditional data science

The COVID-19 pandemic has greatly altered the recent healthcare landscape and has brought about greater minute-to-minute urgency of patient treatment especially in Intensive Care Units (ICUs). This greater urgency for treatment implies a greater need for agility in clinical research, rendering traditional approaches such as Randomized Controlled Trials (RCTs) [1] too slow to be effective. One approach for meeting the agility needs is the use of data science for the development of predictive models to assist in patient treatment. Predictive models can be rapidly and non-invasively developed leveraging existing data and computational tools, and various efforts have been undertaken [2–4]. If successful, predictive models can rapidly process volumes of patient information to assist physicians in making clinical decisions.

However, the development and deployment of predictive models that are useful in clinical environments within short timeframes is challenging. Traditionally, the development and deployment of models employs a sequential process that resembles the *Waterfall* methodology used in software development [5]. Data scientists are given the final requirements of the model by the domain experts, such as physicians. They would begin by searching for training, validation, and test data in sufficient quantities to develop and test the model. The training and

* Corresponding author. CHOC Children's Hospital, Orange, CA 92868, USA. .
E-mail addresses: howard.lei@choc.org (H. Lei), oconnelr@hs.uci.edu (R. O'Connell), lehwerhemuepha@choc.org (L. Ehwerhemuepha), staraman@choc.org (S. Taraman), wfeaster@choc.org (W. Feaster), achang@choc.org (A. Chang).

https://doi.org/10.1016/j.ibmed.2020.100009
Received 9 September 2020; Accepted 18 October 2020

validation data are used to teach the model what to predict, while the test data can be used to evaluate the model. Only after the model meets internal performance requirements would the model be deployed into a real-world setting for domain experts to evaluate and provide feedback. One main disadvantage of this approach is that it prescribes for little collaboration between the day-to-day operations of data scientists and domain experts such as physicians, resulting in data scientists potentially working in isolation for long periods of time. Fig. 1 illustrates this process.

A breakdown of the tasks data scientists typically perform in isolation include data collection, data pre-processing and augmentation, model selection, model hyper-parameter tuning, model training, and model testing. Data pre-processing allows the data to be in a format that's suitable for use by the predictive model. Data augmentation can be used to artificially increase data size, address class imbalances within the data, and make models more robust to potential perturbations within the data. For example, if the input data consist of images, techniques such as translation, scaling, rotation, and adjusting the brightness of images can be used to present more example images for the predictive model to learn. If there is class imbalance within the data, these techniques can also be used to artificially increase the data size for the class with lesser data. Finally, these techniques would allow the model to better handle perturbations of translation, rotation, and brightness variations that may occur in the data. One popular approach for addressing class imbalance and artificially increasing data size is the Synthetic Minority Oversampling Technique (SMOTE) [6]. Note that class imbalance is commonly encountered when working with Electronic Medical Record (EMR) data in healthcare. The class representing the patients with a target condition is typically smaller in size (i.e. with fewer samples) than the class representing patients without the target condition, and this can adversely affect the accuracy of predictive models developed on such data.

Hyper-parameter tuning involves adjusting the parameters used in the model training process [7], where the model is taught how to make predictions given the training data. One example of a hyper-parameter is the number of times – or iterations – the training data is presented to the model to learn. Each iteration is known as an *epoch*. After each epoch the model increases its learning from the training data, and after many epochs the learning is completed. A second example of a hyper-parameter is the percentage of training data that is used by the model in each epoch. The more epochs and the more data presented in each epoch, the better the model learns from the training data. A final example of a hyper-parameter is the learning rate of predictive models. The learning rate represents the step-size that a learning algorithm (i.e. stochastic gradient descent) makes at each step of the model training process. If the learning rate is small, the training may take longer, and the learning algorithm may get stuck in a local minima, which represents a less optimal learned state for the model. If the learning rate is large, the training may be faster but the learning algorithm may "over-step" the global minima, or the optimal learned state. There are various methods for optimizing the learning rate when training models such as CNNs and other kinds of neural networks.

Depending on the amount of training data, the complexity of the data, the number of parameters in the predictive model, and the computing resources, model training can potentially take days to complete. The model would be evaluated against a separate test data to verify if performance meets requirements. If not, some or all previous steps must be repeated until performance becomes acceptable. After the performance is deemed acceptable, the model would be deployed into a real-world environment.

In the end, the process from the conception of the problem to model deployment can take months, and the opportunity for domain experts to evaluate comes only after the deployment of the model. One risk is that after deployment, the model would no longer be relevant if the goals have shifted; another risk is that the model may not meet the performance requirements in a real-world setting. In either situation, time or resources allocated to model development would have been wasted. This can be particularly damaging for data science efforts addressing the COVID-19 pandemic, where rapid development of approaches for detection and diagnoses of symptoms is critical.

2. Challenges for data science in healthcare

In healthcare, the ability to rapidly define goals (i.e. clinically relevant questions) and deploy predictive models that have real-world impact is faced with even more challenges. One challenge is that healthcare data such as Electronic Medical Records (EMR) of patients is inherently complex [8], consisting of a mix of different data types and structures, missing data, and mis-labeled data. The development of predictive models often requires well-structured and well-labeled data; hence, there is a greater need for data exploration, pre-processing and/or filtering when processing EMR data. Furthermore, it may be discovered upon exploration of available training data that the initial clinical questions and goals may not be achievable by predictive models developed using the data. Those questions and goals would need to be refined before model development can proceed.

Furthermore, for predictive models to be useable in a clinical setting, physicians must have confidence that its performance is reliable. Models that perform well under common metrics used by data scientists, such as Area Under the Curve (AUC), does not guarantee that important clinical decisions can be made based on the model [9]. That is because the AUC is a metric that measures model performance across a broad range of sensitivities and specificities of the model. When making important clinical decisions related to patients in the ICU, such as proning versus ventilation, which drugs to use, or whether to administer anti-coagulants, knowing that the model has a high AUC is not as helpful as knowing that a positive decision based on the model has a high chance of being correct. The latter typically occurs when the model is optimized to have

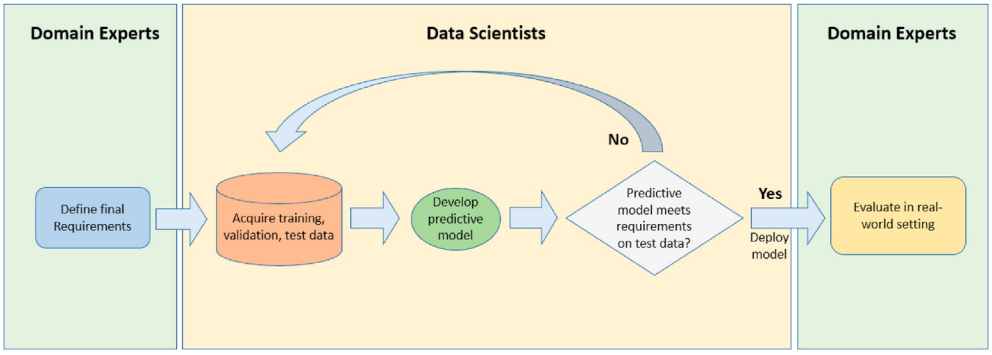

Fig. 1. Traditional approach for development of predictive models, with little collaboration between data scientists and domain experts such as physicians. The model is deployed and evaluated in a real-world setting only after the it has been sufficiently evaluated on the test data.

higher specificity. Some clinical decisions also need to be made within minutes, implying that the model must meet real-time performance standards in order to become the "partner" that can assist physicians in on-the-spot decision making.

The fact that predictive models could lack in performance after being deployed into a clinical setting implies an even greater need for a framework that allows physicians to collaborate with data scientists to continuously monitor model development and performance. Furthermore, the minute-to-minute urgency of treatment needed for the COVID-19 pandemic implies that the lengthy process prescribed by the traditional Waterfall approach – with little communication between data scientists and physicians – is inadequate.

3. Agile data science in healthcare

The agile framework has been traditionally used in software development and has recently been introduced in data science [10]. The framework is an expedient approach that encourages greater velocity towards accomplishing goals. It includes the Scrum and Kanban frameworks, and a hybrid framework called Scrumban [11]. The Scrum framework prescribes consecutive "sprint cycles", with each cycle spanning a few weeks. Within each cycle, team members set and refine goals, produce implementations, and perform a retrospective with stakeholders. New goals and refinements are established for the next sprint cycle. One of the team members also acts as the Scrum Master, who facilitates daily team meetings (called *standups*), and ensures that the team is working towards goals and requirements [12].

The Kanban framework involves breaking down larger tasks into simple, do-able tasks. Each task proceeds through a sequence of well-defined steps from start to finish. Tasks are displayed as cards on a Kanban board, and their positions on the board indicate how much progress has been made [13]. Certain tasks may be "blocked", meaning that something needs to resolve before progress on the task can continue. Fig. 2 shows an example of a Kanban board. One advantage to using a Kanban board is that the set of all necessary tasks, along with progress for each task, is transparent to members of the development team and anyone else who is interested. Overall, the Kanban framework helps bring clarity in tackling larger problems. Domain experts can visualize how the team is tackling the problems, along with what has been accomplished, what is in progress, what still needs to be done, and what needs to resolve before progress can be made.

The Scrumban framework combines Scrum and Kanban. It prescribes multiple sprint cycles, along with defining do-able tasks within each sprint cycle and tracking their progress using a Kanban board. Applying the agile framework to data science in healthcare, we propose the use of the Scrumban framework. For example, each Scrum sprint could contain the following tasks, which can be shown as cards on a Kanban board:

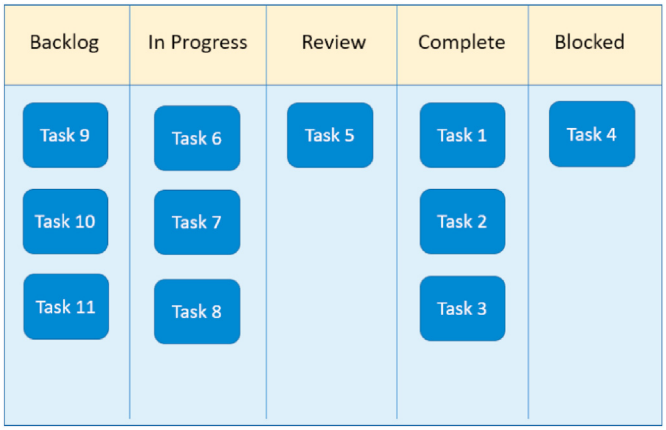

Fig. 2. Example of a Kanban board.

1. Define or refine clinical question(s). *Example: detection of COVID-19 using Chest X-ray images.*
2. Acquire data needed to build predictive models that are aligned with the clinical question(s).
3. Develop a predictive model aligned with the clinical question(s). The goal is to quickly arrive at a model that achieves minimally acceptable performance. This model can be referred to as the Minimal Viable Model (MVM) [14].
4. Deployment of the predictive model into a clinical environment.
5. Data scientists round with physicians to understand clinical usefulness of the model, and repeat steps 1–4 as needed. For example, if the clinical questions are appropriate but the model does not give acceptable performance, then repeat starting from steps 2 or 3. If the clinical questions must be refined, then repeat from step 1. Note that in this step, physicians can also generate new hypotheses if the evaluation of the model yielded new information.

The proposed agile framework is shown in Fig. 3. Unlike the Waterfall approach, the tasks in the agile approach are to be done collaboratively between data scientists and physicians, and we note that the use of cloud-based storage and computing helps by providing a common platform for accessing the data and model(s). Complex problems can be broken down into tasks that can be visualized by both data scientists and physicians, enabling physicians to better understand the work that data scientists must do within each sprint cycle. The framework encourages continuous deployment of predictive models in clinical settings (i.e. such as the ICU), during which time data scientists can round with physicians and receive feedback on the model's performance. The physician's insight or gestalt can be leveraged to determine whether the results of the model are believable [7].

It may be that the predictive model performs well only in certain settings, such as with only certain patient populations across certain periods of time; if so, the clinical questions can be refined or new hypotheses developed at the beginning of the next sprint cycle. The point at which the sprint cycles should end, either because the model has finally become clinically useful or the team needs to pivot towards something different, is determined by the physicians. While the traditional Waterfall approach could take many months for clinically useful models to be developed, the agile approach could take just a fraction of the time depending on the level of collaboration between data scientists and physicians.

4. Implementing agile data science in healthcare

For agile data science to work in the healthcare domain, certain infrastructure must be in place to ensure that sprint cycles can be completed within shorter timeframes. These include the ability to:

1. Rapidly acquire large datasets.
2. Parse and query data in real time.
3. Use established platforms and libraries rather than develop tools de-novo. These platforms and libraries should reside in a cloud framework that allows collaborative efforts to take place.

4.1. Dataset availability

The availability of publicly accessible health information databases for research is increasing despite a multitude of regulatory and financial roadblocks. One such database is the Medical Information Mart for Intensive Care III (MIMIC-III) which contains de-identified data generated by over fifty thousand patients who received care in the ICU at Beth Israel Deaconess Medical Center [15]. The hope is that as researchers adopt the use of MIMIC, new insights, knowledge, and tools from around the world can be generated [16].

Another publicly available database is the eICU Collaborative Research Database, a multi-center collaborative database containing

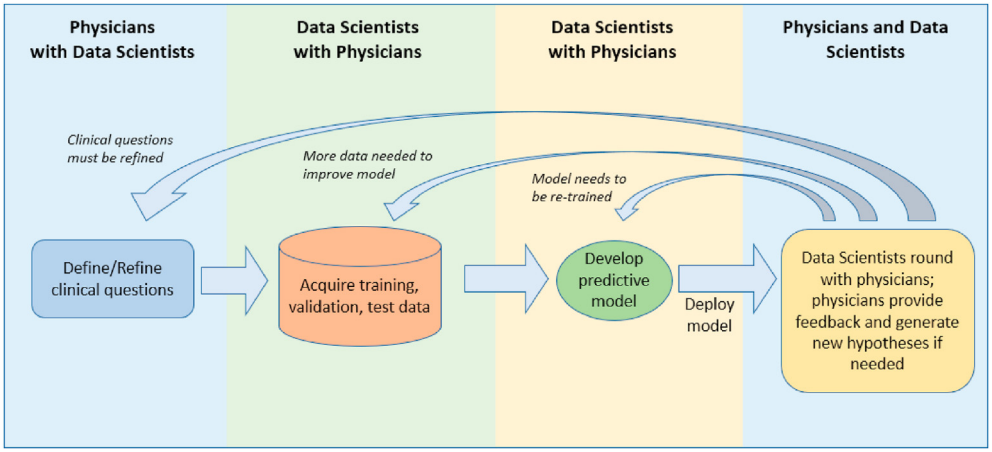

Fig. 3. Proposed agile approach for predictive model development in healthcare. All tasks involve collaboration between data scientists and physicians.

intensive care unit (ICU) data from many hospitals across the United States [17]. Both the MIMIC-III and the eICU databases can be immediately obtained upon registration and completion of training modules. The popularity of these two databases illustrates the potential for large amounts of data to be gathered from hospitals and ICUs around the world and made immediately accessible to researchers. Recently, the five University of California (UC) Clinical and Translational Science Award Institutions including UC Davis, UC Irvine, UC Los Angeles, UC San Diego, and UC San Francisco have collaborated to create a COVID Research Data Set (CORDS), which contains COVID-19 test results across the multiple sites along with patient demographics, past medical history, and lab test results [18]. The Cerner Real-World Data is another research database that contains de-identified data and is freely offered to health systems [19]. Finally, databases for medical imaging studies also exist, such as the Chest X-ray dataset released by the NIH, which contains over 100,000 chest X-ray images [20].

4.2. Cloud storage and computing platforms

Once datasets are obtained, storage and compute power are easily purchased and accessible from an ever-increasing number of vendors. The compute power needed for analyzing large datasets can often be met using cloud computing resources with Amazon Web Services (AWS), Google Cloud Platform (GCP), and Microsoft Azure being the providers of popular cloud services [21–23]. The need for cloud computing tools rests mainly on the availability of specialized elastic compute instances. The elasticity implies that computing resources can be assessed in real time and scaled up or down as needed to balance computing power and cost. Another advantage of a cloud framework is that it allows multiple data scientists and physicians to conveniently collaborate and access the work.

This shift to elastic cloud resources has seen one of the major Electronic Medical Records (EMR) providers, Cerner Corporation [24], develop tools for agile data science that use cloud computing resources as the underlying computing engine. These tools for agile data science often use Jupyter Notebook as the underlying frontend programming interface. Jupyter is an open source computational environment that supports programming frameworks and languages such as Apache Spark][25], Python and R required for processing the data and developing predictive models [26]. Open source deep learning libraries like Keras [27] which enables the rapid development of advanced predictive models such as Convolutional Neural Networks (CNNs) [28], can also be integrated. Finally, the Jupyter Notebook framework supports collaboration amongst multiple individuals, where data scientists and physicians can query data, add and modify code, and visualize results in real time [26].

The availability of the development tools and accessibility of data allow data scientists to rapidly acquire data, query parts of data relevant for addressing clinical questions, and develop predictive models. The outcomes of the model can lead to refinement of the clinical questions, data, or the model itself. The combination of the data scientist, physicians, and agile data science tools will help revolutionize the entire data science process and accelerate discoveries in healthcare and other application domains.

5. Conclusion

Agile data science is quickly becoming a necessity in healthcare, and especially critical given the COVID-19 pandemic. The agile framework prescribes a rapid, continuous-improvement process enabling physicians to understand the work of data scientists and regularly evaluate predictive model performance in clinical settings. Physicians can provide feedback or form new hypotheses for data scientists to implement in the next cycle of the process. This is a departure from the traditional Waterfall approach, with data scientists tackling a sequence of tasks in isolation, without regularly deploying the models in real-world settings and engaging domain experts such as physicians. Given the rapidly shifting healthcare landscape, the goals and requirements for the predictive models may change by the time the model is deployed; this renders the slower, traditional model development approaches unsuitable.

As the agile framework encourages rapid development and deployment of predictive models, it requires data scientists to have easy access to data and the infrastructure needed for model development, deployment, and communication of outcomes. Fortunately, there are now publicly available datasets such as MIMIC-III, and cloud-based infrastructure such as Amazon Web Services (AWS) to achieve this. AWS contains a suite of popular tools such as Jupyter Notebook, Python, and R, allowing data scientists to rapidly upload data, and develop and deploy models with short turn-around time.

Given the increasing amounts of healthcare data, the plethora of clinical questions to address, as well as the minute-to-minute urgency of treating ICU patients given the COVID-19 pandemic, the rapid development of predictive models to address these challenges is more important than ever. We hope that the agile framework can be embraced by increasing numbers of physician and data scientist partnerships, in the process of developing clinically useful models to address these challenges.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

No external funding is provided for this work.

References

[1] Chalmers TC, Smith Jr H, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. Contr. Clin. Trials 1981;2(1):31–49.
[2] Vaishya R, Javaid M, Khan IH, Haleem A. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes, Metab. Syndrome Obes. Clin. Res. Rev. 2020;14(4): 337–9.
[3] Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology 2020; 296(2).
[4] Mei X, Lee H, Diao K, Huang M, Lin B, Liu C, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat. Med. 2020;26:1224–8.
[5] https://www.oxagile.com/article/the-waterfall-model/. [Accessed August 2020].
[6] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 2002;16:321–57.
[7] Liu Y, Chen PC, Krause J, Peng L. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. J. Am. Med. Assoc. 2019;322(18):1806–16.
[8] Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. J. Healthc. Eng. 2018;2018: 4302425. https://doi.org/10.1155/2018/4302425.
[9] Horwitz L. A physician's perspective on machine learning in healthcare. Invited talk presented at Machine Learning for Healthcare (MLHC). MLHC; 2020.
[10] Jurney R. Agile Data Science 2.0. O'Reilly Media, Inc; 2017.
[11] https://www.agilealliance.org/what-is-scrumban/. [Accessed August 2020].
[12] https://www.scrum.org/resources/what-is-scrum. [Accessed August 2020].
[13] https://www.atlassian.com/agile/kanban. [Accessed August 2020].
[14] https://www.linkedin.com/pulse/mvm-minimal-viable-model-farah-chandrima/. [Accessed August 2020].
[15] Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data 2016;3.
[16] Badawi O, Brennan T, Celi LA, Feng M, Ghassemi M, Ippolito A, et al. Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference. JMIR Med Inform 2014;2(2).
[17] Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data 2018;5.
[18] https://www.ucbraid.org/cords. [Accessed August 2020].
[19] https://www.cerner.com/-/media/covid-19/response/2263471793_covid-19-de-identified-data-cohort-access-offer-faq_v1.aspx. [Accessed August 2020].
[20] Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. IEEE CVPR 2017:2097–106.
[21] https://aws.amazon.com/. [Accessed August 2020].
[22] https://cloud.google.com. [Accessed August 2020].
[23] https://azure.microsoft.com/en-us/. [Accessed August 2020].
[24] https://www.cerner.com/ap/en/solutions/data-research. [Accessed August 2020].
[25] https://spark.apache.org/. [Accessed August 2020].
[26] Mendez KM, Pritchard L, Reinke SN, Broadhurst DI. Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing. Metabolomics 2019;15(125).
[27] https://keras.io/. [Accessed August 2020].
[28] Le Cun Y, Bottou L, Bengio Y. Reading Checks With Multilayer Graph Transformer Networks. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 1997.