POSE: Phase II: AutoKeras-OSE - Building an Open-Source AutoML Ecosystem Based on AutoKeras towards Healthcare Applications

Proposal submission on Research.gov ONLY

[DEADLINE: October 21, 2022]Solicitation Link

'**Organization Limit:** Although POSE proposals are expected to be multi-organizational, a single organization must serve as the lead and all other organizations as subawardees. Collaborative proposals arranged as separate submissions from multiple organizations **will not be accepted** in response to this solicitation. Organizations ineligible to submit to this program solicitation **may not receive** subawards; if they are part of the team, their participation is expected to be supported by non-NSF sources.'

- □ Project Summary [One (1) page max].
- □ **Research.gov** documentation
  - □ Collaborators and other affiliations
  - □ Current and pending support
  - □ Bio sketch
  - □ Budget
  - □ Data management
  - □ Equipment and facilities
  - □ Sub-award documentation

JD, please, help me review this checklist

Project Description. (15 pages max)

- □ Context of OSE
- □ Broader Impacts
- □ Ecosystem Growth
- □ Organization and Governance
- □ Community Building
- □ Sustainability

Review Criteria

Phase II proposals will be evaluated on the basis of the following solicitation-specific review criteria:

□ Does the proposal present a convincing case that the OSE will **address an issue of significant societal or national importance** that is not currently being adequately addressed?

□ Does the proposal clearly describe the **long-term vision** for the OSE, including potential partnerships and sustainability?

□ Does the proposal provide convincing evidence that a **substantial user base exists** for the open-source product that will be the subject of the OSE?

□ Does the proposal justify the OSE within the current technological landscape and present a strong case that an OSE is the **best approach for generating impact**?

□ Does the proposal present a clear and comprehensive **description of the ecosystem** within which the OSE will be operating along with **plans for ongoing ecosystem** establishment/**growth** and discovery?

□ Does the proposal present a specific, actionable plan for establishing a **sustainable organizational structure**?

□ Does the proposal present a credible strategy and actionable plan for **building a community of contributors and retaining** contributors?

□ Does the proposal include a clear, detailed **licensing approach** for the open-source product that is the subject of the OSE?

□ Does the proposal clearly describe a **build and test infrastructure**, and procedures to address **quality control and security** of new content?

□ Does the proposal present a clear, actionable evaluation plan to **measure the success** of the OSE with respect to its sustainability goals?

□ Does the proposing team have the **required expertise and experience** to undertake the Phase II activities described in the solicitation?

□ Will NSF support serve as a **critical catalyst** for the establishment and growth of the OSE towards achieving sustainability?

□ Does the proposal include **third-party letters of collaboration** from current users of the open-source technology that is the subject of the OSE?

Project Summary

Overview

The primary goal of this project is to **democratize the development of Automated Machine Learning (AutoML) system to healthcare industries by building up the AutoKeras Open-Source Ecosystem (AutoKeras-OSE).** Specifically, AutoKeras-OSE is built around the widely used AutoML open-source system, namely, AutoKeras, developed by the PI's research group. AutoML has become an increasingly important direction to pursue for both researchers and practitioners. The goal of AutoML is to make ML accessible for domain experts with little ML background knowledge for data-driven discovery and decision-making. We have developed AutoKeras [1], based on Bayesian Optimization and Network Morphism to make neural architecture search very efficient, and empower non-ML domain experts with ML capabilities via simple APIs. AutoKeras is easily downloadable and deployed, using three lines of code, thereby enabling AutoML to be widely applicable. AutoKeras-OSE aims to facilitate a robust, distributed, and inclusive community comprising both healthcare experts as early adopters, and ML developers to seamlessly collaborate via mutual exchange of medical datasets, providing domain knowledge for building use cases, and sharing insights on efficient AutoML model development and deployment.

Intellectual Merit

The key contribution of this project is to enable AutoKeras to be readily used by the healthcare industries. Specifically, we will (1) engage with ML developer community to develop and integrate automated feature extraction and engineering tools in the existing AutoKeras system to identify and extract the most informative features from the input dataset for an end-to-end AutoML pipeline; (2) create several AutoKeras driven healthcare use cases such as Automated Liver Cancer Stratification, and Automated Disease Phenotyping with easy-to-use demonstrations such that domain experts could reuse and deploy the AutoKeras workflow easily in their own infrastructure; (3) build close collaborations and partnerships with Texas Medical Center (TMC), and other medical informatics partners to identify the societal and national needs in healthcare domains; and (4) design community-building programs, and provide support for both developers and users to engage ML and healthcare open-source communities towards creating sustainable and long-term road-map for AutoKeras-OSE development.

Broader Impacts

This project will push forward the frontier of unleashing the great promise of the widely-used open-source AutoML system, AutoKeras, towards healthcare applications, as a ready-to-use tool for democratizing AI techniques to healthcare domain experts without much ML expertise. We will discuss with medical professionals on how to use AutoKeras-OSE, and help them identify the use cases for AutoML within their organizations. Consequently, the results of this project will have an intermediate and strong impact on improving the quality of user experience for healthcare experts, positively impacting the overall value of the machine learning based analytical and information systems, and prompting a more automated platform for emerging and future healthcare applications. The education program will integrate data science, medical informatics, and visualization to train students, and medical professionals without much ML background, and help onboard new contributors. We will further expand our community outreach by incorporating AutoKeras as an integral part of Rice D2K Lab's data-science capstone course projects. PI Hu will continue the existing efforts and reach out to provide both project and research collaboration opportunities to undergraduate, female, underrepresented, and international students.

Keywords: CISE, AutoKeras, Data Preparation Tools, Use Cases, Community Engagement

POSE: Phase II: AutoKeras-OSE - Building an Open-Source AutoML Ecosystem Based on AutoKeras towards Healthcare Applications

1 Introduction

This project aims to **build up the AutoKeras open-source ecosystem (AutoKeras-OSE) for Healthcare applications through user exploration, use case development, community engagement, support and growth in healthcare domains.** Designing machine learning (ML) models to solve real-world problems requires ML expertise, which is time-consuming and labor-intensive. Specifically, to solve the problem with ML model, the pipeline includes serveral steps, including data pre-processing, feature selection, feature engineering, selecting machine learning algorithms, and hyperparameter tuning. Each of the above steps are time-consuming and labor-intensive for the ML expert and daunting for a large section of users without an ML background. This significantly limits the advancement and application of ML for data-driven discovery and decision-making, especially in critical sectors (healthcare), with access to huge data but without much ML expertise.

Automated Machine Learning (AutoML) has become increasingly important for practitioners and researchers to address the above challenges. The AutoML market reached $346.2 million in 2020 and is predicted to reach $14,830.8 million by 2030, demonstrating a CAGR of 45.6% from 2020 to 2030. AutoML aims to make ML more accessible by automatically generating a data analysis pipeline optimized for the data and the target application. Furthermore, automating the development cycle of ML models could yield simpler solutions at a much faster rate with the ability to outperform hand-crafted models.

AutoKeras [1], **led by the PI's research group**, is a leading open-source AutoML software based on Bayesian Optimization and Network Morphism to make the model and hyper-parameter search [2] more efficient and ML accessible to non-ML users. AutoKeras is developed using an open-source deep-learning Python API called Keras. AutoKeras is easily downloadable and deployed, using three lines of code, thereby lowering the barrier for entry and enabling AutoML to be widely used. AutoKeras has become the most popular open-source automated deep learning repository on GitHub (with over 8,500 stars and around 1,400 forks) with a distributed community of 136 contributors spanning over 18 countries and 358 users. The peer-reviewed technical paper introducing AutoKeras published in 25th ACM SIGKDD 2019 has gathered 500+ citations.

1.1 AutoML for Healthcare

Despite the potential of ML in data-driven discovery and decision-making, the ML models are still under-explored in healthcare. Recent efforts have been devoted to applying ML algorithm to solve the healthcare problems. For example, authors in [3, 4] have proposed that ML-based healthcare offers considerable advantages over traditional biostatistical methods for tasks such as risk stratification and survival prediction. Using digital medical imaging data, ML models can acquire diagnostic abilities ranging from physician-level diagnostics [5, 6] to medical visual question answering [7,8]. In general, ML models could improve patient safety [9], improve quality of care [10], and reduce healthcare costs [11]. While ML in healthcare is a very active research topic [12], only 15% of hospitals currently and routinely use ML for even limited purposes [13].

The above lack of deployment and integration of ML-based solutions in clinical settings stems from the following challenges. (1) No single ML model can achieve good performance on all possible healthcare applications that have various data types such as genomics sequences, medical images, Electronic Health Records (EHR), and Electroencephalogram (EEG). This calls for designing custom ML models concerning the actual clinical problem; (2) There is a shortage of

ML expertise, especially in a healthcare setting, with specific technical expertise and knowledge required to develop and optimize ML models. In particular, the collaborative process and interaction between healthcare and ML experts take lots of time and effort from both parties. Hence, there is a need to bridge the gap between ML experts with knowledge of machine learning, and the healthcare experts who have familiarity with clinical data; (3) With varying access to quality datasets, a need for privacy, and distinct compute resources, ML models generally lack interoperability, reproducibility, and interpretability. Therefore, it is crucial to design models to address the above challenges and build confidence in healthcare practitioners, patients, and policy-makers.

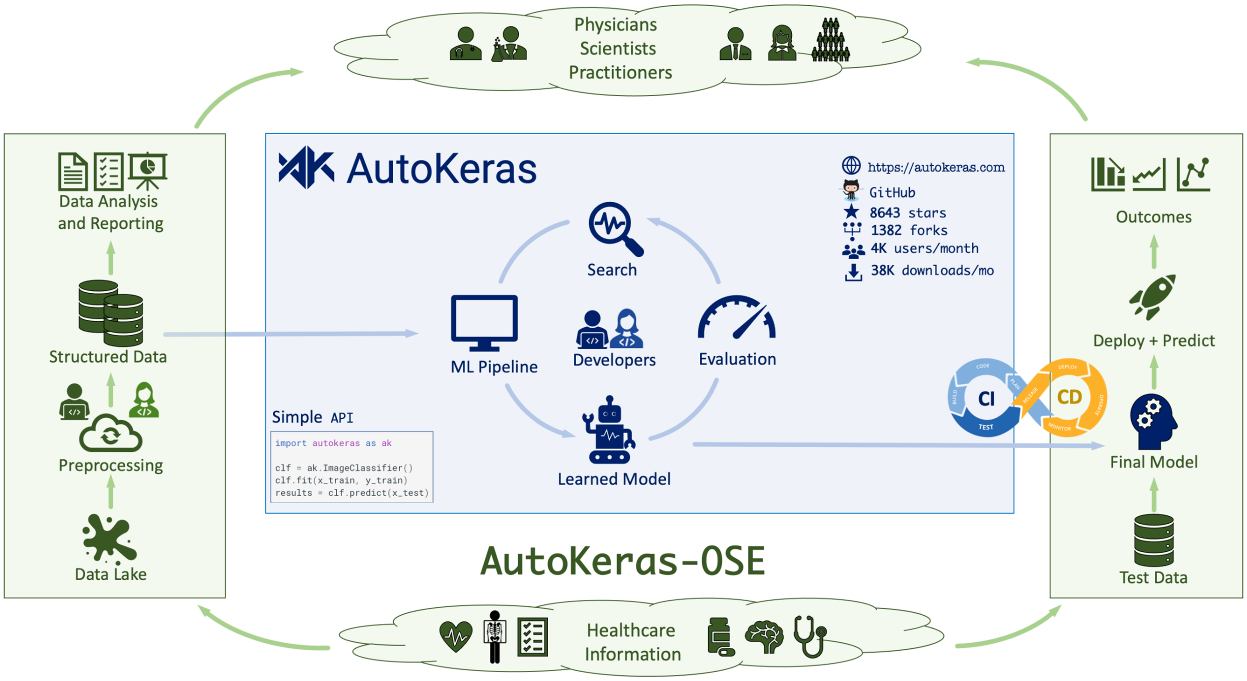
1.2 Need for Open-Source AutoML Ecosystem for Healthcare

The overarching need for an open-source AutoML ecosystem is to build a robust, thriving, and collaborative community comprising ML developers and healthcare users towards lowering the barrier and making ML accessible for accelerating the deployment and integration of ML-based solutions in clinical settings. We discuss the specific needs of the following communities.

Healthcare users comprise research scientists, physicians, and practitioners who are considered domain experts with familiarity with data that arise in their specific target healthcare application and the ability to make clinical decisions. However, they often lack the ML expertise necessary to build and apply advanced ML techniques for data-driven discovery. Healthcare data is being generated at an unprecedented scale having various data modalities. There is a growing need for sound data management with well-curated and labeled datasets. It is challenging for healthcare users to keep pace with the rapid advances made by the ML community in other domains without well-curated healthcare datasets and with a lack of computing and hardware necessary to perform these computations. Moreover, the limited ML models developed through close collaboration between small groups require significant time and communication, and the models may not be reproducible, hindering their applicability in clinical settings.

ML developers are programmers, ML researchers, and data scientists having expertise in ML. They help develop data pre-processing and labeling tools, feature selection/engineering, building optimal AutoML models, model validation and hyperparameter tuning, and deploying models on the target compute infrastructure. However, they lack access to rich clinical and healthcare datasets, which are generally confined to small research groups or organizations. Moreover, ML developers do not have background knowledge in healthcare to understand the domain data and the target problem to build custom AutoML models. As a result, there is a growing need to develop various use cases for applying AutoML in healthcare.

An OSE for AutoML will provide access to highly curated data and compute resources by promoting seamless collaboration between healthcare researchers with datasets and the ML developer community. The ecosystem offers a promising direction for developing novel modeling methods and automated tools for feature engineering/selection and synthetic data generation for different modalities of healthcare data. It will enable the fast creation of optimal ML models customized for target healthcare problems and the available computing resources. Such AutoML solutions will assist in designing and evaluating new assessments, treatments, and methods for enhancing digital clinical trials. Furthermore, the AutoML ecosystem will bring AutoML developers, healthcare researchers, and practitioners together to collaborate on pressing healthcare needs and grow the community seamlessly. This connection will empower the above users without extensive programming and ML experience to build a well-performing model for a target disease, even with a few lines of code. This would significantly enable them to spend more time analyzing the disease pattern and error cases. Moreover, the ecosystem will standardize the dataset collection and labeling process, promote data privacy, build interoperable and reproducible AutoML solutions for given



```
import autokeras as ak

clf = ak.ImageClassifier()
clf.fit(x_train, y_train)
results = clf.predict(x_test)
```

Figure 1: An illustration of the proposed AutoKeras-OSE for healthcare applications. The OSE comprises building a robust and distributed contributor community (denoted as outer green loop) that connects ML experts (developers of automated tools for data pre-processing and AutoML models) with healthcare users (physicians, research scientists, and practitioners) as early adopters of the mature open-source AutoML product, namely, AutoKeras (denoted as inner blue loop).

healthcare problems, and further integrate feature and model interpretation tools for the wide applicability of ML-based models in both healthcare research and practice. This would be achieved by developing various use cases to help developers and users understand the data and apply AutoML models in clinical settings. In light of the above, we propose to build an open-source AutoML ecosystem towards healthcare applications, namely, AutoKeras-OSE which is based on a mature open-source product AutoKeras. Figure 1 illustrates the proposed AutoKeras-OSE.

1.3 Open-Source Ecosystem Thrusts

The long-term goal of our AutoKeras-OSE is to democratize AutoML techniques to people in the healthcare domain. To realize this goal, our detailed AutoKeras-OSE plan includes four thrusts, i.e., identifying potential users, developing and integrating automated feature extraction tools and use cases, supporting users and developers, and building community in healthcare domain.

- □ **Thrust 1: Identifying Potential Users of AutoKeras-OSE in Healthcare Domain.** The thrust aims to train medical professionals and Rice D2K Lab partners on how to use AutoKeras-OSE, and help them identify the use cases for AutoML within their organizations. We establish a partnership with Weill Cornell Medicine, UT Health, and Houston Methodist. We will work closely with these collaborators to develop training resources and curricula that can be used by other healthcare providers around the world.

□ **Thrust 2: Developing and integrating automated feature extraction tools and use-cases for Healthcare data based on AutoKeras-OSE.** The thrust aims to develop automated feature extraction and data synthesis modules for different types of healthcare data which are then integrated within AutoKeras-OSE. Specifically, we design the automated feature extraction tools respectively for following four types of healthcare data (1) Genomic Sequence Data; (2) Electronic Health Record (EHR) Data; (3) Imaging Data; (4) Electroencephalogram (EEG) data. Based on these tools, we also develop several use cases with easy-to-use demonstrations such that domain experts could easily reuse the workflow in their own applications.

□ **Thrust 3: Supporting Users and Developers of AutoKeras-OSE in Healthcare Domain.** The thrust aims to better support the users and developers of AutoKeras-OSE via organizing the following three activates. Specifically, we plan to hold (1) AutoKeras-OSE Consulting which identifies and engages potential users in the inception phase; (2) AutoKeras-OSE Developer Day which encourages developers to design open-source products for healthcare users; (3) AutoKeras-OSE Summit which focuses on the community growth.

□ **Thrust 4: Building AutoKeras-OSE Community in Healthcare Domain.** The thrust aims to build a thriving and sustainable AutoKeras-OSE community. Specifically, we establish three community programs, (1) AutoKeras-OSE membership program aims to organize end-user community, (2) AutoKeras-OSE development partner program aims to recognize and incentivize developers, and (3) AutoKeras-OSE fellow program aims to manage human resources for executing AutoKeras-OSE activities.

2 Context of the Open-Source Ecosystem

2.1 The Guiding Principles and Long-term Vision

The guiding principles of AutoKeras-OSE are: *always learning*, *striving for connection*, and *growing together*. The *always learning* principle emphasizes that AutoKeras-OSE will always learn the expectations and requirements of potential users, so that we can promptly identify and engage potential users. The *striving for connection* principle encourages AutoKeras-OSE to design more open-source products to connect healthcare users and developers for solving real-world problems. The *growing together* principle requires AutoKeras-OSE to create an inclusive environment to support the existing users and developers for growing together. **The long-term vision** for AutoKeras-OSE is to make it easy for everyone to find a ML solution to healthcare problems. We will harness the existing open-source product AutoKeras to democratize AutoML model designing in healthcare based on the guiding principles, and establish a sustainable ecosystem to guarantee AutoML model design is accessible to each healthcare problem-solver.

2.2 Societal Importance

The AutoKeras-OSE will lower the barrier of entry for healthcare users. This is essential for enabling ML capabilities in healthcare for data-driven discovery and decision-making. In addition, it will provide a cost-effective, locally running, and more secure alternative for end-users to design AutoML solutions compared to the existing cloud-based AutoML platforms offered by profit-oriented organizations in industry. This would encourage both the ML developer community and the healthcare domain experts to join and expand the ecosystem into a robust, active, and well-resourced OSE.

2.3 Broader Impact of OSE

The successful outcome of the proposed project will extend the widely used AutoKeras system into healthcare as a ready-to-use tool to democratise AI technology to healthcare practitioners. For healthcare community, we will actively engage with domain experts how to better leverage AutoKeras-OSE in their applications. As a result, this project will have an intermediate and powerful impact on improving the quality of the user experience for medical professionals, positively impact the overall value of machine learning-based analytics and information systems, and lead to a more automated platform for emerging and future medical applications. For the machine learning community, since the dataset in healthcare domain cannot be easily accessed by general public because of the private information of patients, we collaborate with medical professionals to benchmark on biomedical datasets (through our established visiting protocols). In this way, we provide insights for developing novel machine learning algorithm, which may further improve the accessibility of AI technique to both healthcare industries and research communities.

2.4 Automated Machine Learning using AutoKeras

Problem Description. Machine learning has been widely used in many areas, such as computer vision and autonomous driving. However, its technical barrier prohibits the people with limited statistical backgrounds to use ML easily. AutoML is proposed to solve this dilemma by automating the model construction and training. AutoML has become an important research topic, and significant progress has been made in AutoML, such as automated model selection and automated hyperparameter tuning, but there are still several challenges that need to be addressed.

Challenges of AutoML. The challenges for AutoML exist in two aspects. (1) In the context of deep learning, neural architecture search (NAS) aims to automated search fro the best neural network architecture for the given learning task and dataset. Unfortunately, existing NAS algorithms [14, 15] are typically computationally expensive because they require evaluating a huge number of model candidates and need to train each of them in the search space from scratch. (2) The second challenge comes from the deployment of AutoML services. There are several AutoML services available on large cloud computing platforms. Nevertheless, they can be quite limited due to the following reasons: **a)** cost of using cloud services may not be affordable for everyone; **b)** users without a background in computer science may find it difficult to configure cloud-based AutoML using Docker containers and Kubernetes; and **c)** AutoML service providers are honest-but-curious [16], which means they cannot guarantee the security and privacy of customer data.

Solutions. To address the above two challenges, (1) we employed Bayesian optimization for morphing the architecture of a deep neural network while keeping its functionality resulting in efficient NAS designs. (2) We built an open-source software, AutoKeras [1], based on our above technique to make ML accessible to non-ML experts. AutoKeras is easily downloadable and runs locally in contrast to cloud-based AutoML services thereby enabling AutoML to be widely used.

Novelty of AutoKeras. AutoKeras is an open-source AutoML system specifically designed for deep learning tasks. It is challenging to design an easy-to-use and locally deploy-able system. The novelty of AutoKeras lies in the following three aspects. First, AutoKeras has a concise and configurable Application Programming Interface (API) following the classic design of Scikit-Learn API. It is a two-level APIs for both entry level and advanced users. AutoKeras also allows to restore and continue previous searches, search result export and hyperparameter settings for advanced users. Second, AutoKeras runs the program in parallel on GPU and CPU at the same time, which reduces the idle time to improve the efficiency of the search process. Finally, AutoKeras designs a memory estimation function for neural architectures to avoid system crashes due to GPU memory exhaustion. Neural architectures that exceed the upper limit are eliminated.

Evidence of AutoKeras Potential. The AutoML market reached $346.2 million in 2020, and still rapidly grows. The major factors driving the market includes growing demand for efficient fraud detection solutions, personalized product recommendations, intelligent business processes and so on. Many companies have launched their own AutoML products. For example, Amazon Inc. developed Amazon SageMaker Autopilot as an AutoML tool for tabular data. Microsoft Inc. developed their own cloud based AutoML service called Azure AutoML. Paypal Inc. cooperated with H2O.ai to improve the accuracy for fraud detection task to 95% and reduced the model training time to under 2 hours. These AutoML vendors' cloud based services can not avoid the drawbacks in Sec.2.4. AutoKeras , as a mature open-source AutoML software system, has the potential to evolve into an open-source ecosystem of publicly available AutoML services for any ML practitioners and domain experts.

2.5 AutoKeras Details

Current Status. Currently, AutoKeras is an AutoML system under the Keras ecosystem. It is developed by PI Hu's research group (DATA Lab) and jointly maintained with Google. AutoKeras can be easily installed via in Python Package Index (PyPI) or directly downloaded from GitHub. The goal of AutoKeras is to make machine learning accessible for everyone. AutoKeras still keeps expanding its functionalities and is under active maintenance, and its latest version has been updated to 1.0.20.

Development Model. AutoKeras follows the agile development model that relies on the iterative development, intensive communication, and user feedback. Specifically, AutoKeras adopts GitHub Flow, a deployment-centric development model, that allows for continuous, high-speed and secure deployments with simple features and rules. The development process is simple enough and fully automated to support dozens of deployments in a short period of time in real-world development. AutoKeras also follows the semantic versioning rule to solve the dependency hell problem. Given a version number MAJOR.MINOR.PATCH (e.g. 1.0.20), the MAJOR version is incremented for incompatible API changes, the MINOR version is incremented for adding functionality in a backwards compatible manner, and PATCH version is incremented for backwards compatible bug fixes. The users and developers communicate via GitHub where anyone can freely submit their issues and get feedback. The contributors can also provide their code to the project. After aggregating the contribution, AutoKeras updates periodically.

User Base. AutoKeras is developed based on an open-source deep learning Python API, called Keras. With a large community of 136 contributors, AutoKeras has become the most popular automated deep learning repository on GitHub (with over $8,500$ stars and around $1,400$ forks). Over 350 packages have been dependent on AutoKeras. It receives around 38,660 downloads per month and has more than 4,000 monthly active-users, most of whom are students, teachers, researchers and engineers in startups. The peer-reviewed technical paper introducing AutoKeras was published in 25th ACM SIGKDD 2019 and has gathered 500+ citations. AutoKeras is open-sourced under the Apache License 2.0. All the users/developers can copy, modify, update and distribute the source code or the copies of existing software. AutoKeras can be easily installed via PyPI or downloaded from GitHub, and it collaborates well with TensorFlow. AutoKeras contributors are widely distributed around the world spanning over 18 countries including USA, China, India, Canada, France, Switzerland, South Korea, Germany, Spain, Sweden, Japan, Australia, Brazil, Singapore, Portugal, and Bangladesh, etc. In addition, several content creators have independently posted YouTube videos on AutoKeras which has collectively amassed $25,000$+ views. AutoKeras has been also widely adopted in healthcare industry (**see letters of support from existing users**).

3 AutoKeras-OSE: Establishment/Growth

3.1 Identifying Potential Users

As a mature open-source solution, AutoKeras has attracted users from different industries, such as the healthcare industry (**see letters of support from our existing users** from medical schools at UT Health, Baylor, University of Florida, and University of Minnesota). To better describe our users, our team interviewed dozens of supporters who are familiar with AutoKeras to better understand what they value in an open-source product like AutoKeras and how we might best serve them. Healthcare, in particular, was an area that we identified as an important user base. We learned that many healthcare professionals would benefit from the speed and flexibility of AutoKeras , but they face many barriers to adoption. Teaming up with partners and collaborators from the medical schools, we will produce more training resources available for medical professionals interested in AI technology.

Healthcare is an industry with a wide range of data challenges that has the potential to be disrupted by an open-source ecosystem like AutoKeras-OSE, and one that could greatly benefit from the ability to quickly train models on large datasets. As a result of such customer discovery efforts, we established a partnership with Weill Cornell Medicine, UT Health, and Houston Methodist. This high-impact collaboration aims to train medical professionals on how to use AutoKeras-OSE, and help them identify the best use cases for AI technology within their organizations. We'll be working closely with partners and collaborators over the next two years to develop training resources and curricula that can be used by other healthcare providers.

For this proposed ecosystem, we believe that the **user-understanding** process should be a continuous effort rather than a one-time event. Therefore, in this proposal, we also present a continuous user-understanding process that can be applied throughout the lifecycle of AutoKeras-OSE to ensure that designers and developers are always working towards getting users what they need. In particular, we implement the following activities to achieve this goal. (1) Conduct user studies with users at Weill Cornell Medicine, UT Health, and Houston Methodist. These studies will be used to inform an understanding of the needs of specialists in cardiology, population health sciences, and biomedical informatics and how they interact with AutoKeras-OSE. (2) Develop a personas model for people using AutoKeras-OSE that captures their core needs, behaviors, preferences, and attitudes. We strongly believe that our user-understanding efforts sit at the core of our ecosystem as they will help us understand the needs and problems of our potential users moving forward and ensure that AutoKeras-OSE is designed to solve those problems. This close contact with our end user will also inform our understanding of how to best communicate these benefits to potential users so they can see how they would benefit from using AutoKeras .

In the same way, an additional set of potential **regional users of AutoKeras-OSE come from the partners of Rice Center for Transforming Data to Knowledge (D2K Lab)**, in which PI Hu is the director. D2K is a university-wide center for interdisciplinary and experiential data science education and research at Rice. The mission of D2K is empowering students to make an impact in data science through real-world projects in collaboration with medical institutions, industry, government agencies, etc. Since its establishment in 2018, we have already developed long-term connections to over 32 partners in Houston area and beyond, and finished 92 real-world projects. Rice D2K Lab partners comprise many word-leading healthcare institutions. Especially, the Texas Medical Center (TMC) is the world's largest medical center consisting of 54 medicine-related institutions, 21 hospitals, and 8 academic and research institutions. In the framework of Rice D2K Lab, we collaborated with TMC and completed several interdisciplinary research projects about designing ML models to solve healthcare problems, such as biomedical entity recognition in Electronic Health Record (EHR), and the stroke disease predication.

Rice D2K Lab partners will become the optimal partners of AutoKeras-OSE because of three reasons. First, as word-leading healthcare institutions, these partners are at the forefront to tackle the most pressing challenges of healthcare community. Thus they can help AutoKeras-OSE to identify the societal and national needs in healthcare domains. Second, these partners have professional expertise and abundant healthcare data sets. Therefore, they can provide domain knowledge for ML developers and become the early adopters of AutoKeras-OSE to test and validate the open-source products designed for solving healthcare problems. Third, the well-established D2K Lab can facilitate our future collaboration with TMC and other partners, and provides a solid foundation for us to establish AutoKeras-OSE organization and governance, which is essential for AutoKeras-OSE establishment and growth.

3.2 Engaging Contributor Community

Based on user feedback we collected from AutoKeras community, one effective way of engaging contributor community is through development of useful tools and use cases that could enable quick demonstration and easy use of the software. In this subsection, we first introduce several tools that are helpful for maintaining and developing AutoKeras-OSE, and then cover four potential use-cases where AutoKeras-OSE can significantly help.

(A) Integrating/Developing Tools

To better apply AutoKeras-OSE to the healthcare domain, we need corresponding feature engineering and data synthesis modules for integration corresponding to different types of health data. In particular, we organize the integrating/developing tools respectively for (1) Genomic Sequence Data; (2) Electronic Health Record (EHR) Data; (3) Imaging Data; (4) Electroencephalogram (EEG) data. Overall, we aim to build automated tools for feature engineering/selection and synthetic data generation in healthcare domain, achieving a low-barrier entry for healthcare applications with AutoKeras-OSE.

For genomic data, AutoKeras-OSE will automatically generate different types of features from genomic sequences, including biology-informed features and mathematically-focused features. Specifically, AutoKeras-OSE will extract features that fall into the following categories: (a) genome variation features using genome variant calling tools [17]; (b) k-mer based features based on various sizes of k and genome tiles [18,19]; (c) mathematical sequence features based on Fourier mapping, entropy, and complex networks [20]; and (d) DNA shape features, which have been shown to govern DNA-protein interactions [21–23]. To further facilitate AutoKeras-OSE, we also generated synthetic data which could be further used to gain medical insights, which may open the possibilities of many new applications. Specifically, we generate synthetic genomic sequences [24–26] by learning from two aspects: substring and graph. We will generate realistic synthetic components within the sequence with Conditional GAN [27], and combine the generated substrings and generate the realistic synthetic sequence with Graph2Seq [28]. AutoKeras-OSE constructs a graph with the mutual information, where each node is a substring, and each link represents the mutual information between two nodes. Based upon the graph, AutoKeras-OSE then utilizes a Graph2Seq model to capture the higher-order interaction between substrings.

For EHR data, AutoKeras-OSE aims to extract descriptive features, which could be considered as a type of tabular data with categorical and numerical features [29]. Motivated by the success of automated feature augmentation [30], AutoKeras-OSE will adopt a two-stage expanding-selection framework for EHR. In the expanding stage, AutoKeras-OSE will automatically generate three kinds of features, i.e., crossing features such as side-effects of drug-drug-interaction [31], statistical features, and transformed numerical features [32]. In the selection stage, AutoKeras-OSE will select the most useful subset of the generated features to prevent the curse of dimensionality phenomenon [33] and facilitate computational efficiency using SPEC [34] and ReliefF [35]. Further, AutoKeras-OSE will integrate Tabular GAN (TGAN) [36] to synthesize EHR data. TGAN incorporates Gaussian Mixture Model to model the joint distribution of numerical features and categorical features. As a result, TGAN could draw samples from a joint distribution of numerical features and categorical features, where the correlation between them are preserved.

For Medical Imaging data, AutoKeras-OSE aims to explore different types of medical images such as ultrasound, X-ray and CT as illustrated in Figure 2. Images from similar sources could be completely different from each other due to the diversity of machine types and patients' form. To alleviate such issue, AutoKeras-OSE extracts features based on both conventional image processing and learning-based methods. Conventional digital image processing methods such as Local binary pattern [37], SIFT [38,39], and gray level co-occurrence matrix [40] operators will be applied to extract machine-invariant features. Also, AutoKeras-OSE will adopt learning-based methods, i.e., convolutional neural networks (CNN) and U-Net [41, 42], to extract meaningful complex features automatically [43,44]. Besides, AutoKeras-OSE will further synthesize the image data with image augmentation and computer vision techniques. First, AutoKeras-OSE will synthesize new medical images based on data augmentation, such as rotations, cropping, mix-up [45] and their combinations. Second, AutoKeras-OSE will synthesize medical images with deep generative models, such as CycleGAN [46], to synthesize a series of possible variants of images in the real world.

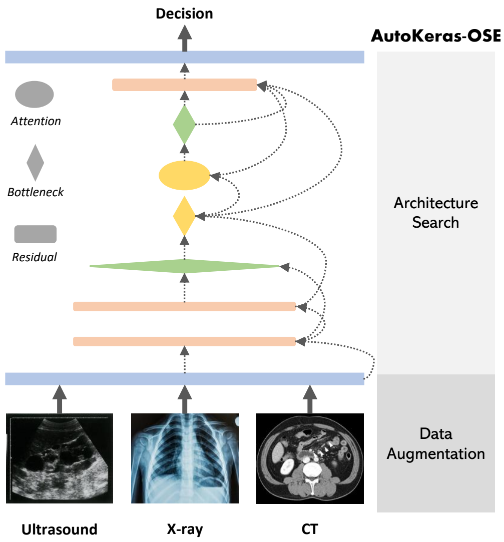


Figure 2: Medical imaging.

For electroencephalogram (EEG) data, AutoKeras-OSE intends to extract effective features which can capture the causal relationships, temporal progression patterns, informative discords, and recurring motifs hidden in the signal. Since EEG signals are essentially one type of time series data, conventional techniques of feature extraction for time series data will be adopted. Specifically, AutoKeras-OSE will extract features via fast fourier transform (FFT), discrete wavelet transform, short-time Fourier transform (STFT), and discrete cosine transformation (DCT) [47, 48]. These methods can extract diverse time domain, frequency domain, and time-frequency domain features from EEG signals [49]. Learning-based approaches will also be adopted to extract features from EEG signals. For example, convolutional neural networks (CNN) and recurrent neural networks (RNN) will be deployed to automatically learn effective features that can be used for various downstream tasks [50–52]. Moreover, AutoKeras-OSE will also generate synthetic EEG data to alleviate the shortage of reliable and subject-specific EEG databy using bandpass filtering, noising, rotation [50], or employing deep generative models (e.g., DCGAN [53, 54]) to synthesize EEG signals from spectrograms.

(B) Developing Use-cases

Automated Liver Cancer Stratification for Precision Medicine: In this task, we aim to build a automated drug discovery pipeline for liver cancer based on AutoKeras-OSE. Liver cancer is life-threatening and quickly becoming the most rapidly increasing cancer in the United States [55]. Patients suffering from liver cancer have to live with poor life qualities and rely on drug and chemotherapy. However, drug development and bedside practice are long and expensive process, while chemotherapy often cause severe side-effects varies by the cancer genotype and patient conditions. Thus, for the studies of liver cancer, gene stratification is an essential approach to implement one of the goals for precision medicine. By introducing AutoML, specific AutoKeras-OSE , high quality features from analyzing genomic sequence data can be extracted. By using the combinatorial search space, AutoKeras-OSE can significantly improve the efficiency and accuracy for clustering the genotype of liver cancers by analyzing the gene stratifications [56].

Automated Lesion Segmentation for Stroke Rehabilitation: This task is to build an automated MRI lesion segmentation system based on AutoKeras-OSE for stroke rehabilitation. Accurate lesion segmentation is the foundation of stroke rehabilitation research [57]. Currently, researchers heavily rely on the manual segmentation results, which is precise but time-consuming and requires lots of neuroanatomical expertise. Moreover, the sizes, shapes, and appearances of lesions in MRI images can vary greatly across different neuroanatomical structures, making the lesion segmentation one of the most challenge task for automated medical imaging algorithms [58]. To solve the problem, AutoKeras-OSE introduces a U-shape based search spaces [41] for the segmentation neural network to enable flexible connection of different-level features based on U-Net.

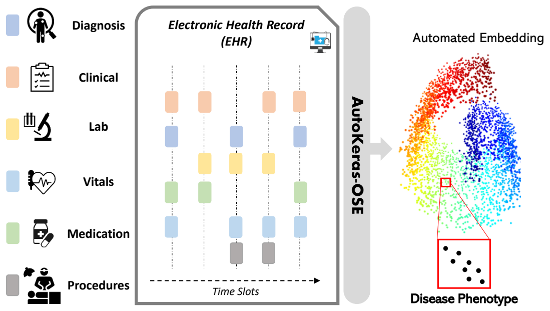


Figure 3: Automated phenotyping.

Automated Disease Phenotyping: In this task, we aim to build an automated disease phenotyping system based on multi-modal electronic health records (EHRs) as illustrated in Figure 3. There is a growing interest in utilizing the EHRs to identify detailed phenotypes for disease diagnosis and research purposes [59]. However, extracting phenotype information from EHRs is not an easy task [60] due to the data fragmentation, multi-modality, and lack of uniform inclusion criteria. In previous works, researchers try to solve this problem by building plenty of rule-based algorithms [61].Rule-based phenotyping algorithms, e.g., Phenotype KnowledgeBase (PheKB) [61] are manually built by researchers and form rules with advanced medical knowledge of the disease [62]. Compared to rule-based algorithms, automated phenotyping with machine learning provides an alternative that could be more generalized and scalable. Specifically, AutoKeras-OSE defines a search space for different modalities, including structured tabular data and unstructured clinic notes. The optimal feature extractors are them selected to capture diease patterns from both tabular and text data. AutoKeras-OSE then leverages a lean-able multi-layer perceptron module to encourage information sharing between two modalities and integrate the representations from both modalities for phenotype prediction. In this way, AutoKeras-OSE can explore complex and implicit patterns for the phenotype prediction, which are often ignored by rule-based methods, improving the robustness and scalability of the EHR phenotyping system.

Automated Sedation Level Monitoring: In this task, we aims to build a automated and robust machine learning system for EEG Monitoring based on AutoKeras-OSE. Monitoring EEG signals is particularly crucial for surgeries in which general anesthesia is necessary, since accurately identifying patients' return of consciousness can help to prevent accidental awareness during surgeries. However, a robust automated machine learning system for EEG monitoring is still missing. Hospitals nowadays employ the Bispectral Index (BIS) monitor to inform anesthetists the sedation level of a patient under general anesthesia [63–67], but the BIS monitor relies on non-personalized statistical measures to make judgements that are often highly inaccurate. The BIS monitor, a device that cannot distinguish different patients or sedative drugs, is not able to produce reliable results, since the patterns in EEG signals can be drastically diverse for different people under distinct sedative medications. Without a reliable monitoring system, inexperienced anesthetists typically tend to overdose sedatives, since accidental awareness can lead to catastrophic consequences such as the post-traumatic stress disorder. However, overdosing sedatives has potential side effects and can lead to symptoms like nausea and sore throat. Therefore, AutoKeras-OSE builds a robust and automated machine learning system that makes subject-specific predictions by training deep neural networks with self-supervised learning strategies, taking advantage of the abundant population data as well as ensuring personalized modeling.

3.3 Supporting Users and Developers

An open-source product in AutoKeras-OSE generally involves three phases: inception, developing, and deployment. To provide comprehensive and dedicates supports to users and developers in all the three phases, we will organize three activities: consulting, developer day, and summit based on the guiding principles of AutoKeras-OSE. In addition, we will extend the AutoKeras *website* to support the users and developers of AutoKeras-OSE.

AutoKeras-OSE Consulting aims to identify and engage potential users in the inception phase, and will be hosted in each quarter. In the consulting event, we will identify the specific requirements and expectations of potential users with respect to open-source products in AutoKeras-OSE . Based on the feedback from potential users, we will elaborate an engaging plan for each potential user, such as offering necessary on-boarding supports, customizing the existing open-source products, and designing new open-source products.

AutoKeras-OSE Developer Day encourages and incentivizes developers to design open-source products for connecting to healthcare users, and will be biannually hosted in each January and July. We will host four sessions to demo the latest open-source products for the four healthcare domains: genomic sequence, EHR, biomedical iamging, and EEG. To incentivize developer to design optimal open-source products, we will organize an open-source product competition for each healthcare domain. Specifically, we will invite healthcare users to submit proposals about ML problems in each domain three months before the developer day, invite worldwide developer to design open-source products for solving the domain problems two months before, and announce competition winners in each domain at the developer day.

AutoKeras-OSE Summit will focuses AutoKeras-OSE community growth, and will be hosted in each October. The summit will consist of a main conference and four workshops. The main conference will invite multiple keynote speakers including healthcare leaders, AutoML researcher, and government officers to discuss any fundamental issues of AutoML and outline the blueprint of AutoML long-term development. The four workshops covers four major topics of AutoKeras-OSE, namely healthcare, innovation, commercialization, and system security, for making a clear and executable road-map of AutoKeras-OSE development. In addition, we will organize several social events for the attendees to seek potential collaborators and expand professional network. Therefore, AutoKeras-OSE will become an thriving and diverse ecosystem.

AutoKeras-OSE Website will be an important infrastructure to support AutoKeras-OSE users and developers. It will provide comprehensive and prompt supports in four aspects: (1) sharing the information and news of AutoKeras-OSE, such as guiding principles, vision and mission statement, organization and governance, and recent events and activities; (2) offering on-boarding and introductory courses for early users to access software platform and development toolkit; (3) establishing a technical forum to help users and developers to exchange ideas and solve issues; and (4) maintaining the existing and developing open-source products.

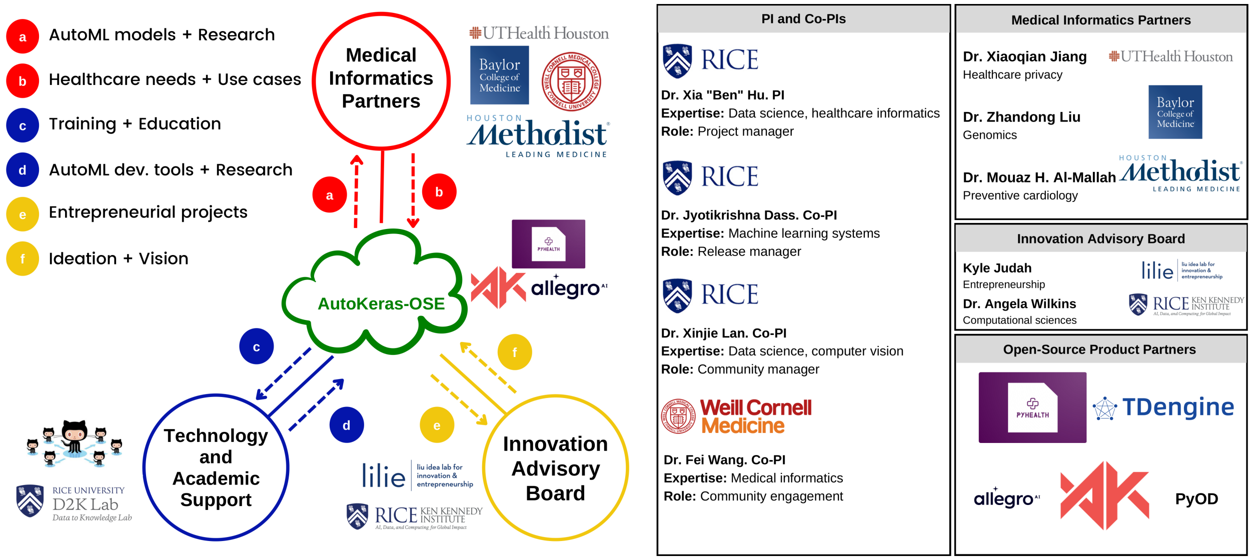

a AutoML models + Research
b Healthcare needs + Use cases
c Training + Education
d AutoML dev. tools + Research
e Entrepreneurial projects
f Ideation + Vision
Medical Informatics Partners
UTHealth Houston
Baylor College of Medicine
Houston Methodist Leading Medicine
AutoKeras-OSE
PyHealth
allegro
Technology and Academic Support
Innovation Advisory Board
Rice University D2K Lab Data to Knowledge Lab
lilie liu idea lab for innovation & entrepreneurship
RICE Ken Kennedy Institute
PI and Co-PIs
RICE
Dr. Xia "Ben" Hu. PI
Expertise: Data science, healthcare informatics
Role: Project manager
RICE
Dr. Jyotikrishna Dass. Co-PI
Expertise: Machine learning systems
Role: Release manager
RICE
Dr. Xinjie Lan. Co-PI
Expertise: Data science, computer vision
Role: Community manager
Weill Cornell Medicine
Dr. Fei Wang. Co-PI
Expertise: Medical informatics
Role: Community engagement
Medical Informatics Partners
Dr. Xiaoqian Jiang
Healthcare privacy
UTHealth Houston
Dr. Zhandong Liu
Genomics
Baylor College of Medicine
Dr. Mouaz H. Al-Mallah
Preventive cardiology
Houston Methodist Leading Medicine
Innovation Advisory Board
Kyle Judah
Entrepreneurship
lilie
Dr. Angela Wilkins
Computational sciences
RICE Ken Kennedy Institute
Open-Source Product Partners
PyHealth
TDengine
allegro
PyOD

Figure 4: An illustration of AutoKeras-OSE organization and governance model comprising medical informatics partners, technology and academic support, and innovation advisory board along with open-source product partners.

4 AutoKeras-OSE: Organization and Governance

Project Team. Figure 4 illustrates the organization and governance model. The team mainly consists of four investigators, with data science and health informatics backgrounds, three medical informatics partners, an innovation advisory board, and five open-source product partners. The PI team comprises expertise from data mining and health informatics (Hu), machine learning systems (Dass), data science applications (Lan), and health data science (Wang). PI Hu is an Associate Professor in the Department of Computer Science at Rice University and leads the Center for Transforming Data to Knowledge (D2K Lab). He has substantial research and management experience in developing automated and interpretable machine learning algorithms to discover meaningful patterns with applications in social informatics and health informatics. Co-PI Dass is a Research Scientist and Program Manager, and Co-PI Lan is an Assistant Teaching Professor, both at the D2K Lab at Rice University. They are managing several active collaborative research projects with medical informatics partners from the Texas Medical Center (TMC). Co-PI Wang is an Associate Professor of health informatics at Cornell University. He has done substantial research in developing machine learning approaches for health data science, including disease phenotyping, disease subtyping, and predictive modeling of clinical risks, etc. In addition to the PI team, the partners are from three major medical schools at TMC, including UT Health, Baylor and Methodist, respectively. The innovation advisors will help us reach out to a wider AI-based healthcare industry with larger impact. The open-source product partners consist several open-source communities that have established the interface with AutoKeras and have been committed to continue the partnership to provide integrated functions in the data science life-cycle.

Licensing. To promote the open-source ecosystem toward broader impacts, we intend to follow the license of AutoKeras for AutoKeras-OSE. Specifically, the Apache-2.0 [68] will be adopted to allow private usages, modification, distribution, commercial purposes, and patent usages at no additional cost or liability to the contributor. The developers who develop the products based on AutoKeras-OSE will have the maximum freedom to license their own products. Meanwhile, the Apache-2.0 will also protect the developers from potential liability of erroneous development that leads to the failures of descendant products. In this way, it will encourage more contributions as well as adoptions from the AutoKeras-OSE.

Continuous Integration / Continuous Deployment. To enable continuous integration of code changes to the repository, we plan to support the following automated scripting. First, we will write integrated and unit tests to cover the majority of the functions in AutoKeras-OSE. Once the committer submits codes to the repository, the auto-testing will be triggered with tailored test cases to ensure the code runs and behaves as desired. Second, we plan to provide building test scripts to ensure that AutoKeras-OSE can be successfully built in different environments. Specifically, the scripts will cover different versions of the programming languages and libraries to ensure that AutoKeras-OSE can serve the majority of the environments. Third, we will design an auto-bot to deal with issues posted by the users or developers. The bot will reply and manage the issues to ensure that the most serious issue will pop out. In this way, the maintainers can easily handle the issues based on the priority and quickly integrate the fix into the main codebase. To support continuous deployment, we plan to follow standard release and deployment procedures of ML projects. First, we will set a regular release cycle of the codebase. In particular, we intend to have a regular release of the code on a weekly or monthly basis. In this way, maintainers and developers can track the evolution of the package. Second, we will follow MLOps, which is a set of practices for the maintenance and deployment of ML models, to ensure the quality of the models, including reproducibility, scalability, etc. To enable the users and developers to easily use AutoKeras-OSE, we will develop infrastructures to support the building and testing. Specifically, we will build and use the servers from D2K Lab to support online demos for users and developers to try AutoML search. We will build a small cluster to handle the computation costs. Specifically, the AutoML search will be conducted in a distributed system, where multiple architecture evaluations will be performed in parallel across servers. To fully guarantee the computational resources for AutoKeras-OSE, we have already got the access to various facilities and equipment provided by Rice University, including Shared Computing Facilities Data Center, Ken Kennedy Institute, Data Center, Networking, Research Data Storage, Collaborative and Archival Systems, OpenStax CNX Open Education Resource, and Rice Supercomputing (NOTS).

Quality, and Security Control. We will maintain and expand our contributing guide with the best practices of contributing code, sharing new ideas and submitting bug reports to AutoKeras-OSE . Specifically, we will document Pull Request Guide, introduce three setup environments such as GitHub Codespaces, VS code and remote-containers, the General Setup, Code Style to format code, and steps for running Tests. We will also maintain a list of Good First Issue, selected for its relative approachability for contributors at the beginner level. For setting up security policy of our public AutoKeras repository, we would leverage the GitHub Advanced Security license that provides code scanning for searching for potential security vulnerabilities and coding errors, secret scanning for detecting secrets (keys and tokens) that have been checked into the repository, and dependency review that shows the full impact of changes to dependencies and see details of any vulnerable versions before merging a pull request. We will also plan to set up Security Control committee for AutoKeras-OSE with access to Security Advisories to discuss, fix, and publish information about security vulnerabilities in AutoKeras.

5 AutoKeras-OSE: Community Building

A thriving and sustainable AutoKeras-OSE community is essential to realize the long-term vision of AutoKeras-OSE . To that end, we establish three community programs: AutoKeras-OSE *membership program* is designed to organize end-users, AutoKeras-OSE *development partner program* aims to recognize and incentivize developers, and AutoKeras-OSE *fellow program* is designed to manage human resources of AutoKeras-OSE . Leveraging the three community programs and the three activities proposed in Section 3.3, we propose a cyclical and enduring strategy for building a thriving and sustainable AutoKeras-OSE community.

AutoKeras-OSE Membership Program aims to support users in a flexible and efficient way. Specifically, we will specify and list all the supports of AutoKeras-OSE, and categorize different supports to three membership levels: premier, strategic, and community members. In general, the basic community members will obtain all necessary supports for open-source product designing, testing, deployment, etc. As a comparison, premier and strategic members can obtain extra supports, such as accelerating product development, reducing technical risk via extra testing and verification, amplifying membership impact via increasing visibility, leadership in AutoKeras-OSE, etc. Potential users can freely choose a membership level based on their specific requirements and the supports they expect to obtain from AutoKeras-OSE .

AutoKeras-OSE Development Partner Program is designed to recognize and incentivize developers to design open source products and collaborate healthcare users to solve problems. This program will assess and evaluate an AutoKeras-OSE developer based on the quantity and quality of open-source products designed by the developer, and recognize outstanding developers who make signification contributions to AutoKeras-OSE. As a return, outstanding developers will have priority to engage in some projects with groundbreaking and innovative healthcare product designing, and their contributions will be advertised to in social media.

AutoKeras-OSE Fellow Program aims to manage human resources to execute the activities of AutoKeras-OSE. Specifically, the fellowship program will appoint, train, and recruit fellows amongst researchers, developers, security experts, communication specialists, financial experts, and students to contribute AutoKeras-OSE, such as training local community members, hosting workshops, organizing social events, and advertising AutoKeras-OSE events in social media. It will empower AutoKeras-OSE user participation and community growth from various domains.

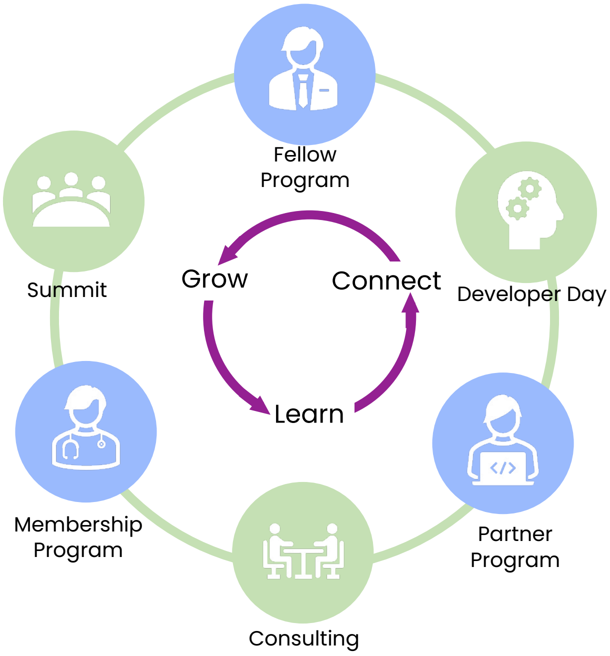


Figure 5: Community building strategy.

The AutoKeras-OSE community building strategy shown in Figure 5 consists of three phases (*learning*, *connecting* and *growing*) based on the three community programs, and the three activities discussed in Section 3.3. In the *learning* phase, we will identify potential users and help them to select an appropriate membership program. In the *connecting* phase, we will connect users to developers for product designing, testing and deployment, and support developers via the developer day event and the development partner program. In the *growing* phase, the summit will invite healthcare users, AutoML researchers and developers to discuss ongoing development plans, existing issues of AutoKeras-OSE, and outline a road-map for future development and maintenance. AutoKeras-OSE fellows will engage and support users and developers in all three community building phases. The three phases form an enduring and cyclical strategy for building a competitive and sustainable AutoKeras-OSE community.

6 AutoKeras-OSE: Sustainability

The sustainability goal of AutoKeras-OSE is to establish a robust and friendly ecosystem. First, robustness is an important aspect of sustainability, because AutoKeras-OSE consists of many open-source software but real-world healthcare data have many exceptions and private information. A sustainable software must tackle these exceptions and avoid data breach and theft to achieve robustness. Second, being friendly plays an significant role in AutoKeras-OSE sustainability: (i) it is easy for a new developer to maintain, extend, and upgrade; and (2) it is easy for a healthcare user to access, understand, and deploy. A sustainable AutoKeras-OSE will foster an inclusive and diverse environment for new developers and healthcare users for easy access.

To achieve the sustainability goal of AutoKeras-OSE, we specify a series of sustainability evaluation metrics with executable actions in the three phases (inception, developing, and deployment) of open-source product. In the inception phase, we will help junior developers to be familiar with product developing requirements and standards via providing open-source product developing tutorial and some introductory courses. A junior developer must finish these courses and obtain the AutoKeras-OSE developing certificate before starting the second phase. In the developing phase, we will provide software and coding checklist to guide open-source product developing. Moreover, we will establish benchmark testing dataset and evaluation metrics based on AutoML theory to assess each product, and create a leaderboard for each healthcare use case to encourage sustainable product development. In the deployment phase, we will form a committee to fully evaluate each open-source product, and help developers to prepare manual book and technical instructions. In addition, we will leverage the commonly used sustainability evaluation metrics, such as the contributors and commits per month, to assess the entire AutoKeras-OSE and assign necessary sources to maintain and support some open-source products.

7 Broader Impacts

The successful outcome of the proposed project will lead to advances in utilizing AtuoML for dealing with challenging health informatics problem. The results of the project will have an immediate and strong impact on improving the usability of machine learning in healthcare applications, positively impacting the overall value of the machine learning based healthcare system, and prompting a more automated and robust platform for emerging and future healthcare challenges. This project will lead to effective, efficient and easy-to-use tools for extracting and engineering features from healthcare data, further advancing healthcare applications, thereby broadly impacting the healthcare field. This high-impact collaboration with medical schools enable training medical professionals and students on how to use to solve their problems. Thus the project has the potential to reach both the healthcare industry and academic community.

This project will play an integral part in educating and training students. We will work closely with Weill Cornell Medicine, UT Health, and Baylor to develop training resources and curricula that can be used by other healthcare providers. The research will also be tightly integrated with related courses on data science and health informatics at Rice University and the other partnering medical schools. Moreover, we also expand our community outreach and educational activities by incorporating AutoKeras as an integral part of Rice D2K Lab's data-science capstone course projects. Moreover, PI Hu has a track record in providing research opportunities to undergraduate, female, underrepresented, and international students, and would like to continue these efforts throughout the duration of the project. This course project will educate students on how to approach their own applications with AI techniques, and thus potentially improve student success rate and decision-making, which may later be published to a general education system.

Letters of collaboration. Minimum 3 and maximum 5. [Ben]

'These letters of collaboration must be from current users or contributors (who are not directly related to the proposing team) of the open-source product that is the subject of the proposed OSE. Each letter writer should clearly describe how they have contributed and will continue to contribute to the development of an OSE including the technical advancements enabled by these contributions and the value proposition associated with the product.'

- □ Done - Letter of collaboration from Xiaoqian Jiang at UT Health
- □ Done - Letter of collaboration from Zhandong Liu at Baylor
- □ Done - Letter of collaboration from Kyle Judal
- □ Done - Letter Zhe He from UFL
- □ Done - Letter from Rui Zhang at UMN

A list of Project Personnel, Collaborators, and Partner Institutions

□ Xia (Ben) Hu; Rice University; PI

□ Jyotikrishna Dass; Rice University; Co-PI

□ Xinjie Lan; Rice University; Co-PI

□ Fei Wang; Cornell University; Co-PI

□ Xiaoqian Jiang; University of Texas Health Science Center at Houston; Unpaid Collaborator

□ Zhandong Liu; Baylor College of Medicine, Texas Children's Hospital; Unpaid Collaborator

□ Mouaz H. Al-Mallah; Houston Methodist; Unpaid Collaborator

□ Kyle Judah; Liu Idea Lab for Innovation and Entrepreneurship; Unpaid Collaborator

□ Angela Wilkins; Rice University; Unpaid Collaborator

□ Rui Zhang; Department of Surgery, University of Minnesota; Unpaid Collaborator

□ Zhe He; Florida State University; Unpaid Collaborator

Data Management Plan

This document is to introduce the data management plan for our AutoKeras-OSE project, which mainly includes two parts, i.e., the security plan and data handling plan. Specifically, The security plans contains three parts, i.e., accessibility, adoption, and contribution. In the data handling plan, the data can be grouped into the following five categories: websites, data, publications, software, and curriculum materials. Below we first introduce the details for the security plan.

Security Plan. Security Plan contains three parts, i.e., accessibility, adoption, and contribution. For the accessibility, AutoKeras-OSE will extend the AutoKeras website for data access and guarantee data security via establishing a secure password-protected portal. The web server is maintained by Rice University, and the access to this secure portal will only be shared by the verified users and developers of AutoKeras-OSE. For the adoption, AutoKeras-OSE adopts the in-house policy for security consideration, i.e., the security problem is directly solved by the people who governs AutoKeras-OSE internally. For the contribution, we will engage community contributor to inspect and report the potential security problem of AutoKeras-OSE.

Data Handling Plan. Note that we do not expect to directly access any sensitive data in the proposed project, instead, we mainly work with publicly available datasets. The data involved in AutoKeras-OSE can be roughly divided into the following four categories.

1. Project Website: We will create a website for AutoKeras-OSE to facilitate data and software access, sharing, and reuse. It serves as a convenient entry point to improve the efficiency of communication between students, researchers and collaborators. It will provide links to various outcomes of the project including links to a data repository, a publication list, and a software library, as well as course websites, curriculum materials, and links to external sources. The general public can conveniently reproduce the results on the website using AutoKeras-OSE.

2. Data Repository: Datasets that are collected following standard strategies of will be archived in a dedicated repository as parts of AutoKeras-OSE and made publicly available. Mainly we will collect data from social media platforms and health care systems. The repository will be an extension to our existing data repository, algorithms, and evaluation results. In the published datasets, we will ensure that the identities of individuals are not revealed and cannot be recovered. We want to make data available to expedite the advancement of research, but also maximally protect any user information by state-of-the art anonymization techniques that can help identify users. Only anonymized versions of the data will be retained in AutoKeras-OSE such that cannot be used to identify specific individuals. Open standards as defined by the research community will be used to format the data prior to sharing.

3. Software Library: Novel algorithms and computational tool-kits will be developed as parts of AutoKeras-OSE and extensive experiments will be conducted to evaluate the new kind of algorithms by comparing with the state-of-the-art algorithms. Public access and support will be provided to assist research efforts at large in developing open-source algorithms to help demonstrate and understand the strengths and weaknesses of new methods. To better illustrate our understandings and findings, some demonstration software systems will be provided. Useful open-source tools related to Data Mining and Machine Learning will be used in our regularly offered courses at Rice University. The software will be licensed under the open source standard.

4. Curriculum Materials: The curricular materials consist of class notes, slides, and hands-on projects based on AutoKeras-OSE. The materials prepared by the PI for education purposes such as class notes and slides will be made publicly available. The data related to students' performance will only be returned to the individual student to protect their privacy. All assessment results, which are publicly shared via reports, tutorials, or any other similar venues, will provide an aggregate analysis to prevent identification of students in order to protect their privacy.

SUPP: Budget [Ben: It is being process by Rick]

’**10** Phase II proposals of up to **$1,500,000** in total budget, with duration of **up to two years**. Can include support for Salary, Setup costs, and Mandatory training. **Refer Page 6** of solicitation.’

References

[1] Haifeng Jin, Qingquan Song, and Xia Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1946–1956, 2019.

[2] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

[3] Kee Yuan Ngiam and Wei Khor. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, 20(5):e262–e273, 2019.

[4] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[5] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, and Josep Malvehy. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, 2020.

[6] Colin Jacobs and Bram van Ginneken. Google's lung cancer ai: a promising tool that needs further validation. *Nature Reviews Clinical Oncology*, 16(9):532–533, 2019.

[7] Qi Wu, Peng Wang, Xin Wang, Xiaodong He, and Wenwu Zhu. Medical vqa. In *Visual Question Answering*, pages 165–176. Springer, 2022.

[8] Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12):5456, 2021.

[9] Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, and Jerry Kim. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760, 2018.

[10] Huiying Liang, Brian Y Tsui, Hao Ni, Carolina Valentim, Sally L Baxter, Guangjian Liu, Wenjia Cai, Daniel S Kermany, Xin Sun, and Jiancong Chen. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature medicine*, 25(3):433–438, 2019.

[11] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health affairs*, 33(7):1123–1131, 2014.

[12] Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *Jama*, 319(13):1317–1318, 2018.

[13] Xueqiang Zeng and Gang Luo. Progressive sampling-based bayesian optimization for efficient and automatic machine learning model selection. *Health information science and systems*, 5(1):1–21, 2017.

[14] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

[15] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[16] Qi Chai and Guang Gong. Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers. In *2012 IEEE international conference on communications (ICC)*, pages 917–922. IEEE, 2012.

[17] Geraldine A Van der Auwera and Brian D O'Connor. *Genomics in the cloud: using Docker, GATK, and WDL in Terra*. O'Reilly Media, 2020.

[18] Christopher Fletez-Brant, Dongwon Lee, Andrew S McCallion, and Michael A Beer. kmer-svm: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucleic acids research*, 41(W1):W544–W556, 2013.

[19] Mahmoud Ghandi, Dongwon Lee, Morteza Mohammad-Noori, and Michael A Beer. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology*, 10(7):e1003711, 2014.

[20] Robson P Bonidia, Lucas DH Sampaio, Douglas S Domingues, Alexandre R Paschoal, Fabrício M Lopes, André CPLF de Carvalho, and Danilo S Sanches. Feature extraction approaches for biological sequences: a comparative study of mathematical features. *Briefings in Bioinformatics*, 22(5):bbab011, 2021.

[21] Tianyin Zhou, Lin Yang, Yan Lu, Iris Dror, Ana Carolina Dantas Machado, Tahereh Ghane, Rosa Di Felice, and Remo Rohs. Dnashape: a method for the high-throughput prediction of dna structural features on a genomic scale. *Nucleic acids research*, 41(W1):W56–W62, 2013.

[22] Tianyin Zhou, Ning Shen, Lin Yang, Namiko Abe, John Horton, Richard S Mann, Harmen J Bussemaker, Raluca Gordân, and Remo Rohs. Quantitative modeling of transcription factor binding specificities using dna shape. *Proceedings of the National Academy of Sciences*, 112(15):4654–4659, 2015.

[23] Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. Dnashaper: an r/bioconductor package for dna shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, 2016.

[24] Zhen Lin, Michael Hewett, and Russ B Altman. Using binning to maintain confidentiality of medical data. In *Proceedings of the AMIA Symposium*, page 454. American Medical Informatics Association, 2002.

[25] Bradley A Malin. Protecting genomic sequence anonymity with generalization lattices. *Methods of information in medicine*, 44(05):687–692, 2005.

[26] Jessamyn Dahmen and Diane Cook. Synsys: A synthetic data generation system for healthcare applications. *Sensors*, 19(5):1181, 2019.

[27] Kaleb E Smith and Anthony O Smith. Conditional gan for timeseries generation. *arXiv preprint arXiv:2006.16477*, 2020.

[28] Kun Xu, Lingfei Wu, Zhiguo Wang, Yansong Feng, Michael Witbrock, and Vadim Sheinin. Graph2seq: Graph to sequence learning with attention-based neural networks. *arXiv preprint arXiv:1804.00823*, 2018.

[29] Derek Snow. Deltapy: A framework for tabular data augmentation in python. *Available at SSRN 3582219*, 2020.

[30] Yuanfei Luo, Mengshuo Wang, Hao Zhou, Quanming Yao, Wei-Wei Tu, Yuqiang Chen, Wenyuan Dai, and Qiang Yang. Autocross: Automatic feature crossing for tabular data in real-world applications. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1936–1945, 2019.

[31] David L McCollum, Jerica L Greene, and Darren K McGuire. Severe sinus bradycardia after initiation of bupropion therapy: a probable drug-drug interaction with metoprolol. *Cardiovascular drugs and therapy*, 18(4):329–330, 2004.

[32] Fatemeh Nargesian, Horst Samulowitz, Udayan Khurana, Elias B Khalil, and Deepak S Turaga. Learning feature engineering for classification. In *Ijcai*, pages 2529–2535, 2017.

[33] Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks*, pages 758–770. Springer, 2005.

[34] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[35] Marko Robnik-Šikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1):23–69, 2003.

[36] Lei Xu and Kalyan Veeramachaneni. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264*, 2018.

[37] Devrim Unay, Ahmet Ekin, Mujdat Cetin, Radu Jasinschi, and Aytul Ercil. Robustness of local binary patterns in brain mr image analysis. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2098–2101. IEEE, 2007.

[38] Warren Cheung and Ghassan Hamarneh. N-sift: N-dimensional scale invariant feature transform for matching medical images. In *2007 4th IEEE international symposium on biomedical imaging: from nano to macro*, pages 720–723. IEEE, 2007.

[39] Stéphane Allaire, John J Kim, Stephen L Breen, David A Jaffray, and Vladimir Pekar. Full orientation invariance and improved feature selectivity of 3d sift with application to medical image analysis. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–8. IEEE, 2008.

[40] Mellisa Pratiwi, Jeklin Harefa, and Sakka Nanda. Mammograms classification using gray-level co-occurrence matrix and radial basis function neural network. *Procedia Computer Science*, 59:83–91, 2015.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[42] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018.

[43] Junyu Chen and Eric C Frey. Medical image segmentation via unsupervised convolutional neural network. *arXiv preprint arXiv:2001.10155*, 2020.

[44] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *European Conference on Computer Vision*, pages 762–780. Springer, 2020.

[45] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[46] Jue Jiang, Yu-Chi Hu, Neelam Tyagi, Pengpeng Zhang, Andreas Rimner, Joseph O Deasy, and Harini Veeraraghavan. Cross-modality (ct-mri) prior augmented deep learning for robust lung tumor segmentation from small mr datasets. *Medical physics*, 46(10):4392–4404, 2019.

[47] Amjed S Al-Fahoum and Ausilah A Al-Fraihat. Methods of eeg signal features extraction using linear analysis in frequency and time-frequency domains. *International Scholarly Research Notices*, 2014, 2014.

[48] Murugappn Murugappan, Mohamed Rizon, Ramachandran Nagarajan, S Yaacob, I Zunaidi, and D Hazry. Eeg feature extraction for classifying emotions using fcm and fkm. *International journal of Computers and Communications*, 1(2):21–25, 2007.

[49] Robert Jenke, Angelika Peer, and Martin Buss. Feature extraction and selection for emotion recognition from eeg. *IEEE Transactions on Affective computing*, 5(3):327–339, 2014.

[50] Chaoqi Yang, Danica Xiao, M Brandon Westover, and Jimeng Sun. Self-supervised eeg representation learning for automatic sleep staging. *arXiv preprint arXiv:2110.15278*, 2021.

[51] Nihal Fatma Güler, Elif Derya Übeyli, and Inan Güler. Recurrent neural networks employing lyapunov exponents for eeg signals classification. *Expert systems with applications*, 29(3):506–514, 2005.

[52] Arthur Petrosian, Danil Prokhorov, Richard Homan, Richard Dasheiff, and Donald Wunsch II. Recurrent neural network based prediction of epileptic seizures in intra-and extracranial eeg. *Neurocomputing*, 30(1-4):201–218, 2000.

[53] Nik Khadijah Nik Aznan, Amir Atapour-Abarghouei, Stephen Bonner, Jason D Connolly, Noura Al Moubayed, and Toby P Breckon. Simulating brain signals: Creating synthetic eeg data via neural-based generative models for improved ssvep classification. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[54] Khansa Rasheed, Junaid Qadir, Terence J O'Brien, Levin Kuhlmann, and Adeel Razi. A generative model to synthesize eeg data for epileptic seizure prediction. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:2322–2332, 2021.

[55] Jones PD Morris C Ransdell JM Kwon D Brown CP Kobetz EN Pinheiro PS, Callahan KE. Liver cancer: A leading cause of cancer death in the united states and the role of the 1945-1965 birth cohort by ethnicity. *JHEP*, 1, 2019.

[56] L Kotthoff. Algorithm selection for combinatorial search problems: A survey. *Data Mining and Constraint Programming*, 10101, 2016.

[57] Peter Langhorne, Julie Bernhardt, and Gert Kwakkel. Stroke rehabilitation. *The Lancet*, 377(9778):1693–1702, 2011.

[58] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

[59] Katherine I Morley, Joshua Wallace, Spiros C Denaxas, Ross J Hunter, Riyaz S Patel, Pablo Perel, Anoop D Shah, Adam D Timmis, Richard J Schilling, and Harry Hemingway. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. *PloS one*, 9(11):e110900, 2014.

[60] Jyotishman Pathak, Abel N Kho, and Joshua C Denny. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*, 20(e2):e206–e211, 2013.

[61] Jacqueline C Kirby, Peter Speltz, Luke V Rasmussen, Melissa Basford, Omri Gottesman, Peggy L Peissig, Jennifer A Pacheco, Gerard Tromp, Jyotishman Pathak, and David S Carrell. Phekb: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association*, 23(6):1046–1052, 2016.

[62] Jessica K De Freitas, Kipp W Johnson, Eddye Golden, Girish N Nadkarni, Joel T Dudley, Erwin P Bottinger, Benjamin S Glicksberg, and Riccardo Miotto. Phe2vec: Automated disease phenotyping based on unsupervised embeddings from electronic health records. *Patterns*, 2(9):100337, 2021.

[63] Sara Afshar and Reza Boostani. A two-stage deep learning scheme to estimate depth of anesthesia from eeg signals. In *2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME)*, pages 7–12. IEEE, 2020.

[64] Yue Gu, Zhenhu Liang, and Satoshi Hagihira. Use of multiple eeg features and artificial neural network to monitor the depth of anesthesia. *Sensors*, 19(11):2499, 2019.

[65] Ravichandra Madanu, Farhan Rahman, Maysam F Abbod, Shou-Zen Fan, and Jiann-Shing Shieh. Depth of anesthesia prediction via eeg signals using convolutional neural network and ensemble empirical mode decomposition. 2021.

[66] Sara Afshar, Reza Boostani, and Saeid Sanei. A combinatorial deep learning structure for precise depth of anesthesia estimation from eeg signals. *IEEE Journal of Biomedical and Health Informatics*, 25(9):3408–3415, 2021.

[67] Ejay Nsugbe and Stephanie Connelly. Multiscale depth of anaesthesia prediction for surgery using frontal cortex electroencephalography. *Healthcare Technology Letters*, 2022.

[68] Andrew Sinclair. License profile: Apache license, version 2.0. *IFOSS L. Rev.*, 2:107, 2010.