**Introduction**

Phase equilibrium (flash) calculations is a long-standing issue in thermodynamics due to the sudden nature of the change and its computational cost. Using method such as Successive substitution method to correlate and predict vapor-liquid equilibrium takes a lot of time due to long-term iteration. Due to the high time consumption in forward model, there is an increasing effort on substituting forward model simulations by using faster surrogate models based on machine learning. Yu Li et al (2019) implemented deep model in flash calculation and show good result

Machine learning methods, have had a great impact in different fields, including, self-driving cars, identification of images and due to its black-box nature and flexibility are being used in different fields to try to solve long standing issues with great success in some areas. Due to far exceeding other model’s accuracy and broad versatility in all kinds of region, deep leaning method has become one of the most popular models in the past few years.

Yu et al 2019, shows that, using the variations of DNN and show good result in flash calculations, even by cross comparation to other model, such as Newton’s method and the Sparse Grids Method.

However, in 2022, Ravid Shwartz-Ziv (et al) use another family of model named Tree ensemble models, which shows the great power of classical machine learning algorithms. The Tree ensemble model family (such as gradient-boosted decision trees (GBDT)) still have dominant status in processing tabulate data (Chen T et al,2016). When neural network meets reality problem, It is often accompanied by a range of thorny problems such as lack of locality, data sparsity and mixed feature types. While at the same time, the tree-ensemble algorithms require much less tuning and less computation cost to fit data. In addition, Ravid shows that ensemble of deep models and XGBoost performs better on these datasets than single model alone

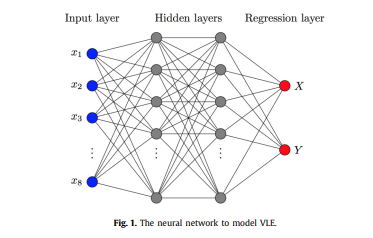
The main goal of this study is to find a model to accelerate flash calculation, while at the same time, we need to compare the existing models mentioned in the two paper above and try to Evaluate in different dimensions. While also we can also build library to help us implement this method in further study.

Yu Li (et al,2019) use deep learning method trying to replace this process.

**Literature Review**

Vapor-liquid equilibrium (VLE) calculation is of great importance in modeling and simulating flows with multiphase and multicomponent. In the last few decades, people raised numerous experiences and obtain tremendous data trying to enhance model performance with equations of state method (EOS). The EOS method, simulate real process in experience like combination and interaction between binary mixtures, can produce accuracy result in the end. However, in most cases, such method needs a number of iterations to reach the equilibrium statement, which often limits the practicability in real case. In order to deal with this problem, several methods were raised such as Newton’s method and Sparse grids method. Both of them did well in their age.

In 2012, the breakthrough of AlexNet draw people’s attention into deep learning model. The model, with amazing ability, showing its strength in all kinds of region. It can not only carry out well in original computer vision field, setting new record in image recognition and language translation, but can also been noticed to have great performance in many other problems such as protein binding affinity prediction [5], enzyme function prediction [17], structure super-resolution reconstruction and modeling brain circuits (Yu Li et al,2019). Yu Li(2019) studied how to implement artificial neural networks in VLE calculation and compare it with traditional algorithms. He found out that deep learning does have advantage in VLE alculation and even outperform some “elder brother” algorithms in processing such tabulate data such as support vector machine (SVM)



The ANN Li used, similar to other neural network, takes several variables as input, such as critical pressure (P c), critical temperature (T c) and process it by linear combination layer and activation function, then it is transfered into next layer and so on, the math formular will look like this:

yi = fi(Wi ∗ ai + bi),

where Wi is the weight; bi is the bias; and fi is the activation functions of the i-th layer

Meanwhile, He also use multiple technique like Weight decay and Dropout method to gain better function of model

It seems that deep learning has become silver bullet in all kinds of region, however in 2022, a new paper writed by Ravid Shwartz-Ziv, Reestablished leadership of machine learning method in processing tabulate data by using Tree ensemble models (such as XGBoost). The author also admits that some deep networks do have strength in some situation as well. In this paper, He collected all kinds of model (such as DNF-Net, 1D-CNN) and all kinds of dataset from various place (such as Microsoft, Shrutime) on the Internet to systematically compare the performance between models and between datasets. In the end of the paper, the author shows that XGboost perform better compare with other model and in addition he concludes that Ensemble model of XGBoost and deep models can score better results than XGBoost alone. The detail of study that Ravid have down will be introduced below:

**Differentiable trees**: Family member of old brothers in processing tabulate data and have excellent performance. It takes advantage of good property in Quadratic function and fix some training problems in classical decision trees. In addition, it supports parallal programming and approximate algorithms, which is really good in our practical study.

**Attention-based models:** Attention-based model has been widely used since 2017. Recent works have suggested that not only inter-sample attention, but intra-sample attention is also important.

**Model ensemble**: modeling ensemble is a widely accepted way to generating better result and reduce fluctuation in different dataset. Consequently, ensembles tend to increase it stability and get better results

And also, these models have been used in various dataset (like Gesture Phase, Gas Concentrations, Microsoft (MSLR) and so on). Several analysis methods have also been used like Statistical significance test. Due to his Scientific and Systematic work, the article gives an instructiveness suggestion in this region and in our future study.

**Description of Problem and** **Objectives**

-------to be continue

**Progress to Date and Future Plan**

|  |  |
| --- | --- |
| **Dates** | **Task** |
| 1 Jun – 10 Jun | Literature review and build Generate data library. |
| 10 Jun – 20 Jun | Form test module/log module/simple model (GBDT, ANN) to build more powerful and stable library |
| 20 Jun – 30 Jun | Test the simple model and try to fix it into larger dataset and build system that used to find best hyperparameter of model |
| 1 July-20 July | Try to implement more model with best hyperparameter and collect data to analysis it  1 use more complexed model, like 1D-CNN, transformer  2 use model trained by more information, like Physical informed model |
| 20 July- 25 July | visualize and analysis data, review what need to be improved |
| After 25July | ## Haven’t figure out, besides writing report, what can I done? |

**Reference**

Yu Li, Tao Zhang, Shuyu Sun, Xin Gao, Accelerating flash calculation through deep learning methods, *Journal of Computational Physics*, Volume 394, 2019, Pages 153-165, ISSN 0021-9991. <https://doi.org/10.1016/j.jcp.2019.05.028>.

T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

Ravid Shwartz-Ziv, Amitai Armon, Tabular data: Deep learning is not all you need, Information Fusion, Volume 81,2022, Pages 84-90, ISSN 1566-2535,

<https://doi.org/10.1016/j.inffus.2021.11.011>.