**Drug Discovery Hackathon**:

**PS: DDT2 – 14:** **A computational pipeline to predict Drug Induced Liver Injury (DILI).**

***Team: Mastercard (AI Garage)***

***Registration ID:* 16986334**

**Approach**:

There are 4 unique pillars which go into the model-

1. UGRNN
2. Graph Kernel
3. Auto-Encoder on PaDel Features
4. Cat-Boost Classification Model
5. **UGRNN**: Treat each molecule as a graph & apply LSTM on it to learn the embedded pattern.
6. **Graph-Kernel**: It compares within class & between class graphs & learns embeddings of each molecule graph in a comparative manner.
7. **Auto-Encoder on PaDel Features**: It compresses the required information of PaDel features from ~1444 D to 8 D
8. Now we’ve 3 distinct sets of embeddings which are created in a mutually exclusive manner & capture different areas of information.  
   Thus, we train a **Cat-Boost** Model on top of all these embeddings to learn the final mapping and make predictions.

**Applicability Domain** - We are only considering Molecules where all the atoms are connected i.e. there is no floating Atom.

**Our Benchmark Results –**

We did a comparative study on Dataset of Published Paper - **Deep Learning for Drug-Induced Liver Injury** (<https://pubs.acs.org/doi/10.1021/acs.jcim.5b00238>) and we are getting better results in terms of Accuracy, Sensitivity, Specificity, MCC and G-Mean.

On Test Set

|  |  |  |
| --- | --- | --- |
| **Metric** | **Our Results** | **Paper Results** |
| Accuracy | **87.6** | 86.9 |
| Sensitivity | **82.7** | 82.5 |
| Specificity | **95** | 92.9 |
| AUC-ROC | 0.93 | 0.95 |
| MCC | **0.763** | 0.746 |
| G-Mean | **88.3** | 87.5 |

On Train Set

|  |  |  |
| --- | --- | --- |
| **Metric** | **Our Results** | **Paper Results** |
| Accuracy | **90.9** | 88.4 |
| Sensitivity | **90.7** | 89.9 |
| Specificity | **91.3** | 87 |
| AUC-ROC | **0.99** | - |
| MCC | **0.819** | 0.771 |
| G-Mean | **89.1** | 88.3 |