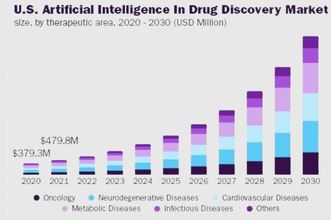
Prediction of Antimicrobial Peptides Based on ML

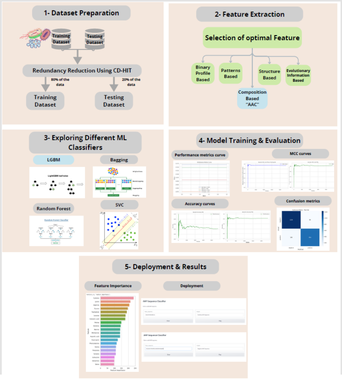
Motivation

- Antimicrobial resistance is an urgent and global health problem as existing drugs are becoming ineffective against the treatment of antimicrobial infections. In this bioinformatics project, we aim to build a machine learning model to predict antimicrobial peptides that help in the development of novel Antimicrobial Drugs. It is killing at least 1.27 million people worldwide and associated with nearly 5 million deaths in 2019. In the U.S., more than 2.8 million antimicrobial-resistant infections occur each year, 1 in 5 people who died from AMR was a child under 5 years old, often from previously treatable infections.
- Here the importance of artificial intelligence appears, as the global AI in drug discovery market size reached $0.6 billion in 2022 and is projected to grow at a CAGR of 45.7% to reach $4.0 billion by 2027, according to a new report by MarketsandMarkets:



Objective

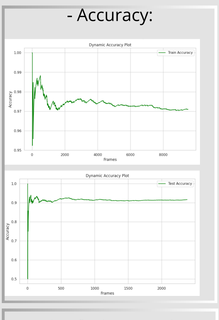
- Building a machine learning model to predict antimicrobial peptides project aims to predict the Antimicrobial peptides “AMPs” by using a pre-trained ML model on some important AMPs features instead of analyzing them manually in laboratories “ which takes a lot of time and effort”, this in turn helps in monitoring the Antimicrobial Resistance “AMR” and using it in drug discovery & development process.



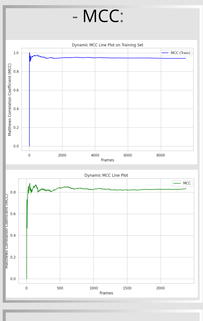
Results and analysis

- After applying the methods used in the project “starting from collecting the data step, to training the LGBM model”, it’s time of model evaluation. For this step, we used many evaluation metrics and highlighted features as shown as follows:

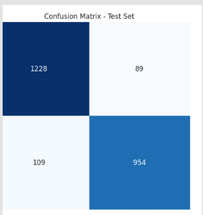
- Accuracy:



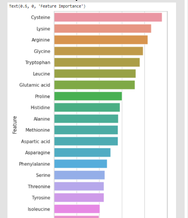
- MCC:



- Confusion Metrics:



Feature Importance result:

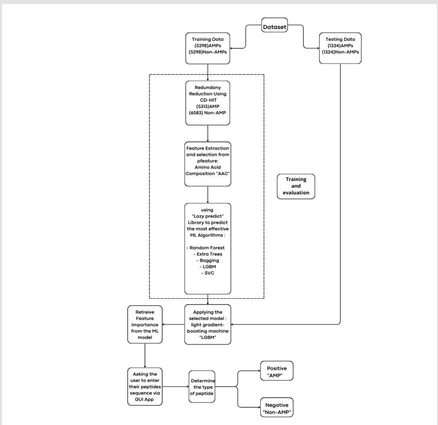


Team members

Salah Gamal Abdelkhabir	**4211099**
Ahmed Ebrahim Gabr	**4211102**
Seham Samy Ezz-Eldeen	**4211225**
Omnia Khamis Khalil	**4211028**
Ahmed Hamdy Eldakroury	**4211126**
Ahmed Mohamed Ali	**4211095**
Kareem Hamed Abdelghaffar	**4211068**
Shahd Lotfy Hamed	**4211192**

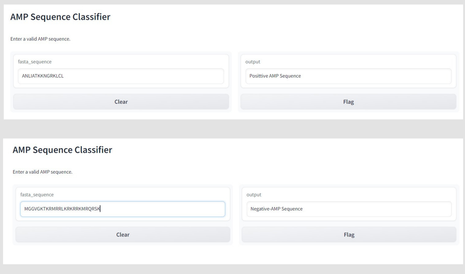
Materials and methods

- **Data Preparation:**
  We employed CD-HIT for redundancy reduction in training and testing datasets with clustering threshold = 0.99.
- **Feature Extraction:**
  Applied Composition-Based approaches (Amino Acid Composition) which extracts 20 Amino Acids from the sequence using Pfeature.
- **ML Classifiers:**
  After testing 30 classifiers with lazypredict, actually testing the top ten classifiers, We ultimately chose LightGBM (LGBM) as our final model.
- **Model Training & Evaluation:**
  We splitted dataset in ratio 80% training and 20% testing data, then we evaluated via accuracy curves, confusion matrix, and MCC curves.
- **Deployment:**
  We had made API with gradio to make user able to enter the sequence, then knows if it efficient or not.



Implementation

- After applying our methods:
  - Data collection
  - Data preprocessing
  - Feature extraction
  - Fit suitable model
- we had to deploy the model to enable user to interact by entering a sequence for peptides then result appeared if it is efficient AMP or not.



Future work

- In the future, we will have to make some enhancements related to the project itself like more data and try to get it from real labs and apply more efficient methods like ensemble learning and data augmentation and deploy with a mobile or web application to be more usable.
- And other enhancements by developing the project in a new larger idea like drug discovery.

Conclusion

- Our project “Prediction of Antimicrobial Peptides Using Machine Learning” integrates powerful data science techniques to classify peptides into positive or negative antimicrobial peptides (AMPs) through model creation and deployment.
- All that will advance studying immune system, monitoring Antimicrobial Resistance (AMR) and is scalable in next stages.