# An Explainable Machine Learning Platform for Antimicrobial Resistance Prediction and Resistance Gene Identification

**MIT PRIMES Andrew Zhang, Roxbury Latin School**

**Mentor: Dr. Gil Alterovitz**

# Introduction

## Antimicrobial Resistance - a Global Health Crisis

- **Antimicrobial resistance (AMR)** - bacteria or microbes evolve to resist certain antibiotics
- Worldwide, 700,000 people die annually due to AMR infections
- It is estimated that there will be 10 million deaths worldwide each year by 2050 if the trend continues
- World Health Organization (WHO) identifies AMR as one of the biggest threats to global health

# Problem

- A Challenge of treating AMR infection is long diagnosis time: two days to even a month with current culture-based diagnosis
- Current practice in clinics for AMR treatment is broad-spectrum antibiotic therapy
  - contributes to more AMR, oftentimes ineffective
- Rapid diagnosis of AMR in clinics urgently needed to save patient lives and prevent outbreaks
- Furthermore, diagnosis ought to be explainable, as this would be helpful for understanding the mechanisms behind resistance

# Proposal

**Using genomic data, Deep Convolutional Neural Networks (DCNNs) can predict AMR Accurately, and Support Vector Machine (SVM) can find the mutations that cause the resistance**

- AMR phenotype is determined by genomic data, a strain's chromosomal resistance mutations
- DCNN: advanced machine learning algorithm
  - Highly successful in image classification
  - High speed, accuracy
- Convert genomic data to image format
  - Analogous to image classification
- SVM intrinsically explainable - usable as a surrogate to DCNNs to pinpoint the genes and mutations that cause AMR

# Data Collection

## Genomic Data Retrieval

- Genomic mutations of 149 strains of *Mycobacterium tuberculosis* is collected from international hospitals
- Data on their resistance to pyrazinamide (PZA) is obtained from culture-based testing
- Genomic data is in Variant Call Format (VCF)
- Build a parser to gather all mutations from VCF. A snapshot of a strain's mutations are as below

| POSITION | MUTATION |
|---|---|
| 698 | G->A |
| 1977 | A->G |
| 4013 | T->C |

# Method

- **Convert Mutations To Genomic Images**
  - The union of all mutations from all strains form a list, with each entry having the position, the reference base, and alternate base.
  - For each strain, build an array initialized with all 0's
    - if it has a mutation that belongs to the union, the array element is set to 1
  - We pad each array with zeros so it can be reshaped an n x n array – "**Genomic Image**", similar in format to a traditional image represented with pixels
- **Represent AMR phenotypes as one-hot arrays**
  - Resistant strains are mapped to label [0, 1]; susceptible strains to label [1, 0]

# Method

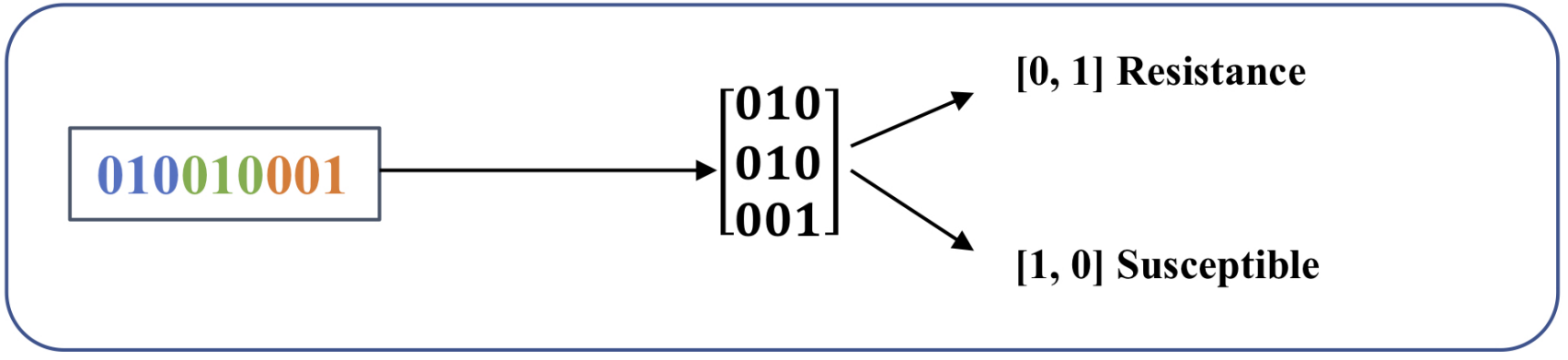
## Converting Mutations to Feature Array

**Union of mutations for all strains**

| | Position | Mutation | Gene |
|---|---|---|---|
| **1** | 7361 | C/T | Rv0006 |
| **2** | 7362 | G/C | Rv0006 |
| **3** | 7539 | A/G | Rv0006 |
| **4** | 2289215 | T/G | Rv2043c |
| **5** | 2289221 | T/G | Rv2043c |
| **6** | 2289227 | T/G | Rv2043c |
| **7** | 4247351 | G/A | Rv3795 |
| **8** | 4247577 | A/C | Rv3795 |
| **9** | 4247607 | A/G | Rv3795 |

**Mutations of strain BTB13-128**

| | Position | Mutation | Feature array |
|---|---|---|---|
| **1** | N/A | N/A | 0 |
| **2** | 7362 | G/C | 1 |
| **3** | N/A | N/A | 0 |
| **4** | N/A | N/A | 0 |
| **5** | 2289221 | T/G | 1 |
| **6** | N/A | N/A | 0 |
| **7** | N/A | N/A | 0 |
| **8** | N/A | N/A | 0 |
| **9** | 4247607 | A/G | 1 |



# Method

## Build a DCNN Model for AMR Prediction

- Four convolutional layers–filter size 4x4, and five filters per layer, as shown in Figure 2.
- At the last convolutional layer (Conv4 Layer), a large stride is used to reduce the dimension before a fully connected(FC) layer.
- No pooling layer is used, though they are widely used in image classification
  - Pooling would cause info loss in Genomic Image resulting in lower prediction accuracy
  - Use strides in convolution layer for dimension reduction instead of using pooling layers

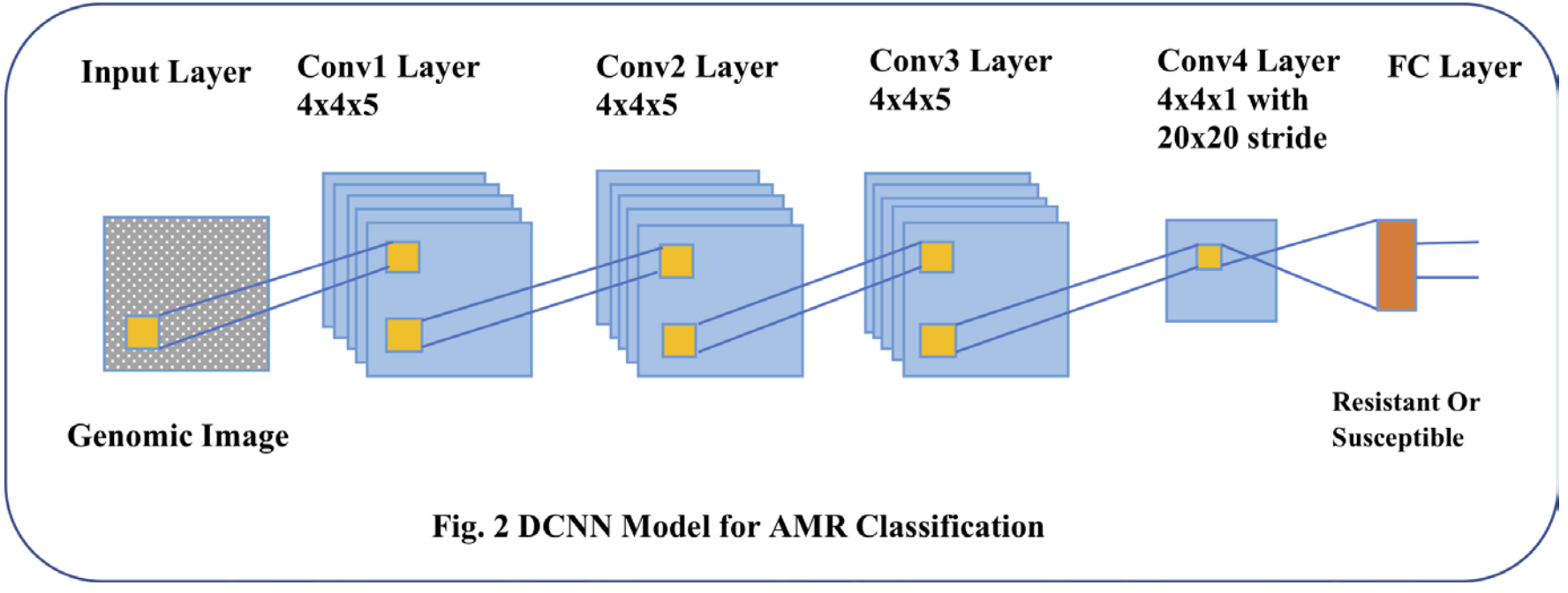



Fig. 2 DCNN Model for AMR Classification

# Method

## Build a Support Vector Machine Model to Identify Genes/Mutations that Harbor Resistance

- DCNN achieves high accuracy, but does not identify genes and mutations that harbor resistance
  - In medical applications, important to explain diagnosis
  - Explainable diagnosis help understand root cause and provide better therapies
- A SVM model is built as a surrogate to DCNN model
  - SVM is inherently explainable
  - The weights of the trained SVM indicate the importance of a feature
- A SVM model finds a hyperplane that divides the inputs to two classes with the maximum margin between the two
  - The resistant strain's input array is mapped to -1
  - The susceptible strain's input array is mapped to 1

# Method

## A Recursive Feature Addition Algorithm uses SVM to find resistance gene list, PZA_genes

1. From all the resistance genes, pick one gene a time to find the gene, gene_0, that produce the most accurate prediction using the SVM model. Add gene_0 to PZA_genes, so PZA_genes = {gene_0}.
2. Take another gene from the remaining genes one by one, gene_i, train the SVM model with gene_i and the genes in PZA_genes list. Pick the gene_i that produces the most accurate prediction, and is better than the result without gene_i by at least a margin, delta. If gene_i is found, add it to PZA_genes, so PZA_genes = {gene_0, …, gene_i}. If gene_i can not be found, stop.
3. Repeat step 2

# Procedure

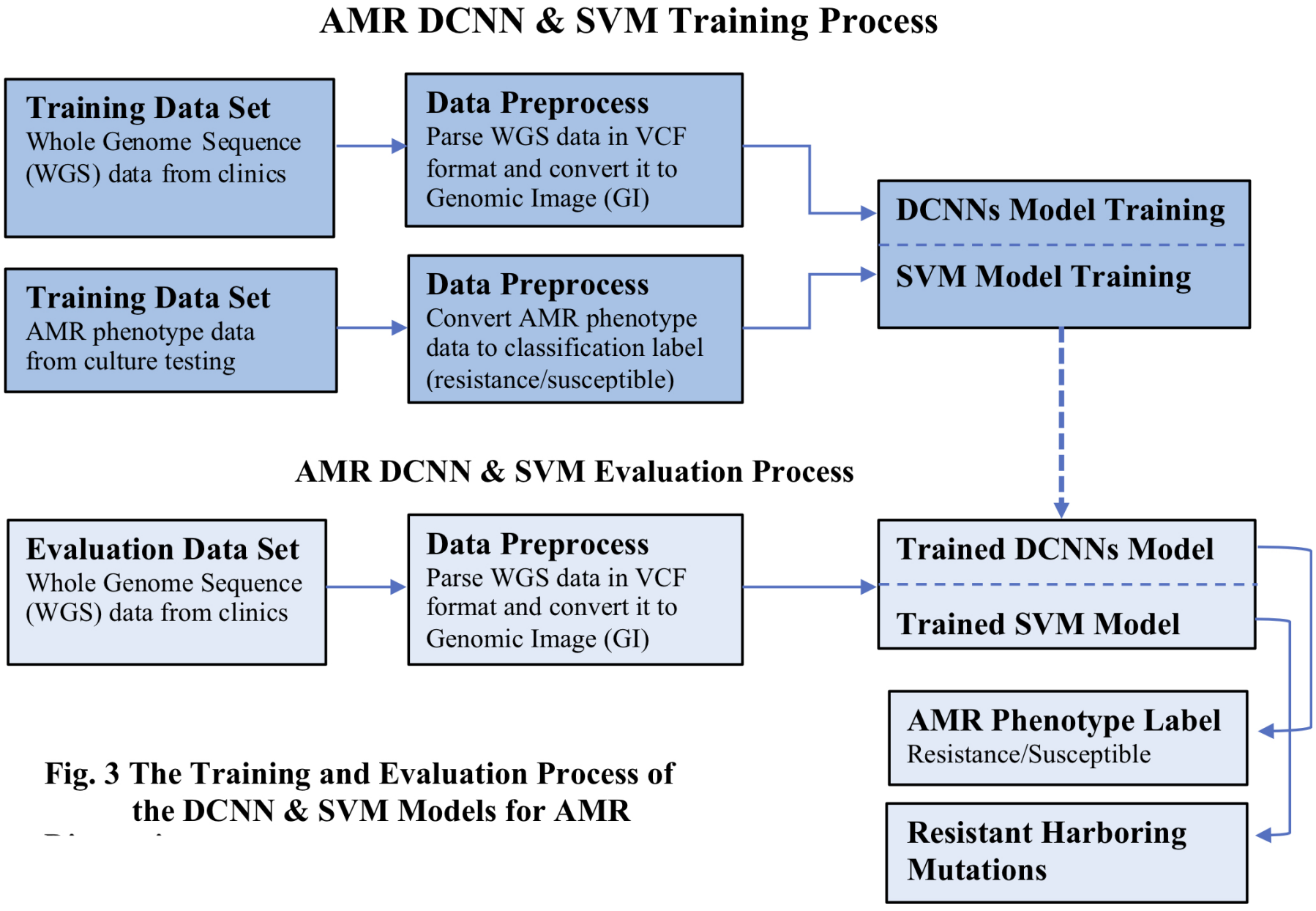



**Fig. 3 The Training and Evaluation Process of the DCNN & SVM Models for AMR**

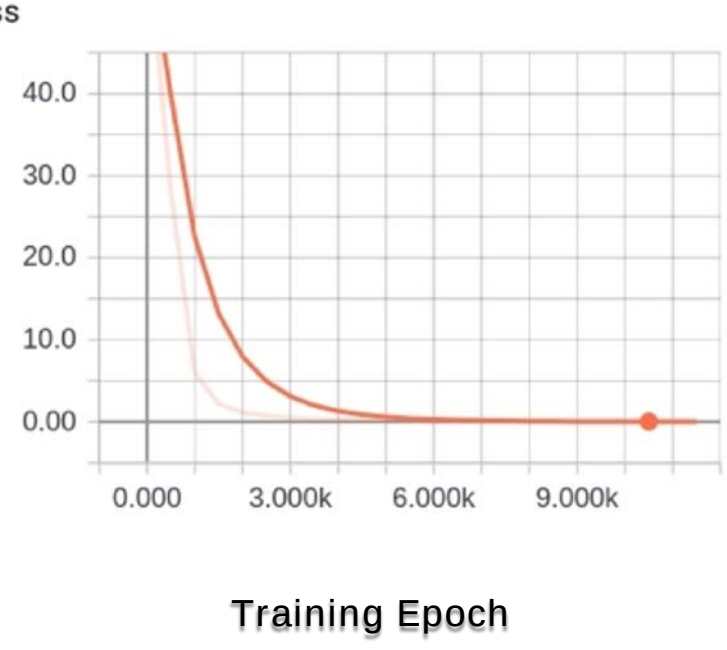
# Result

## Test Case 1: Prediction of M. Tuberculosis Resistant to Pyrazinamide (PZA)

120 strains for training and 29 strains for evaluation

### Training Result:

- Loss function is reduced to 0 in 9000 epochs
- Loss function is defined as the quadratic of the difference between model output and the actual label
- Loss reaching 0 means model output and label obtained from bacterial database fully match




### Evaluation Result:

- Only 2 out of 29 strains has wrong prediction – 93.1% accuracy.
- Prediction time is less than 1 second

# Result

## Test Case 2: Find Resistance Genes Using the SVM Model and the Recursive Feature Addition Algorithm

| Genes | Accuracy with training data | Accuracy with evaluation data | Total Mutations | Unique Mutations |
|---|---|---|---|---|
| pncA | 87% | 64% | 77 | 42 |
| pncA+embB | 85% | 78% | 199 | 79 |
| pncA+embB+gyrA | 89% | 80% | 704 | 104 |
| embB+gyrA | 79% | 72% | 627 | 62 |

- The algorithm starts with the gene *pncA*, a well-known PZA resistance gene
- It recursively adds one gene a time to get better accuracy
- The algorithm finds two other genes, *embB* and *gyrA*, which harbor PZA resistance
- The *embB* gene was also found to be correlated with PZA resistance in literature with different dataset, and *gyrA* is a novel gene

# Result

**Test Case 3: Find Mutations that Cause the Resistance**

| Position/Gene | Mutation | Resistant Strains | Susceptible Strains |
|---|---|---|---|
| 4247607/pncA | A->G | 15 | 7 |
| 4247609/pncA | G->A | 12 | 5 |
| 2289076/embB | G->C | 5 | 1 |
| 2288969/embB | A->C | 4 | 0 |
| 7581/gyrA | G->T | 2 | 0 |

- Using the weights of the trained linear SVM model, I have found 9 mutations that contribute the most to resistance (5 of them in the table).
- As PZA resistance is mapped to “-1”, and susceptible is mapped to label “1”, the mutation of a strain corresponding to the most negative weight, is the mutation that contributes the most to the resistance.

# Conclusion

- AMR diagnosis is tackled as a Genomic Image Classification problem using DCNN
- Diagnosis accuracy of 93.1% in less than a second
- A surrogate SVM model is built to identify resistance genes and mutations
- Create a Recursive Feature Addition algorithm that identifies three genes that harbor PZA resistance
- SVM pinpoints 9 mutations that caused the PZA resistance
- The rapid diagnosis could guide doctors to use the right antibiotics to save patients' life and prevent the outbreak of the deadly AMR infections.
- Identifying the genes and mutations helps understanding the biological mechanism of the resistance and combating it

# Future work

- Understand the biological mechanisms by which mutations cause the PZA resistance
- Run the ML platform side-by-side with culture-based diagnosis in a hospital lab

# References


1. The Economist, Antibiotic use is rapidly increasing in developing countries, 2018. https://www.economist.com/graphic-detail/2018/04/02/antibiotic-use-is-rapidly-increasing-in-developing-countries
2. The Centers for Disease Control and Prevention, Antibiotic/Antimicrobial Resistance, 2018. https://www.cdc.gov/drugresistance/index.html
3. United Nations meeting on antimicrobial resistance. Bull World Health Organ. 2016;94(9):638-9.
4. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13(9):601-12.
5. Lin RY, Nuruzzaman F, Shah SN (2009) Incidence and impact of adverse effects to antibiotics in hospitalized adults with pneumonia. J Hosp Med 4:E7–E15
6. Costelloe C, Metcalfe C, Lovering A, Mant D, Hay AD (2010) Effect of antibiotic prescribing in primary care on antimicrobial resistance in individual patients: systematic review and meta-analysis. BMJ 340:c2096–c2096
7. Sheen P, Requena D, Gushiken E, Gilman RH, Antiparra R, Lucero B, et al. A multiple genome analysis of Mycobacterium tuberculosis reveals specific novel genes and mutations associated with pyrazinamide resistance. BMC Genomics. 2017;18(1):769.
8. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y, Dillon N, et al. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. Nat Commun. 2018;9(1):4306.
9. Santerre JW, Davis JJ, Xia F, Stevens R, Machine learning for antimicrobial resistance. arXiv preprint arXiv:1607.01224. 2016. https://arxiv.org/pdf/1607.01224.pdf
10. Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, et al. PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. Infect Immun. 2011;79(11):4286-98.
11. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. J Mach Learn Res 12:2825–2830
12. Rawat W, Wang Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Comput. 2017;29(9):2352-449.
13. Krizhevsky A, Sutskever I, Hinton G., ImageNet Classification with Deep Convolutional Neural Networks, Advances in neural information processing systems. 2012
14. Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Explainable AI: The new 42? Paper presented at the CD-MAKE 2018, 27-30 Aug 2018, Hamburg, Germany
15. Adadi A., Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI), IEEE Access. 2018;6:52138–52160.
16. Molnar C., Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2019.
17. The Variant Call Format Specification, VCFv4.3, 2019, https://samtools.github.io/hts-specs/VCFv4.3.pdf
18. Abadi M, Barham P, Chen J, et al (2016) TensorFlow: A System for Large-Scale Machine Learning. 265–283
19. Kingma DP, Ba J (2014) Adam: A Method for Stochastic Optimization.
20. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS Med. 2009;6(2):e2.

# Acknowledgement


- Many thanks to Dr. Gil Alterovitz and Dr. Insung Na for their guidance on this project and providing the clinical data
- Thank you to the PRIMES program for giving me this opportunity
- Thanks to Dr. Gerovitch and Prof. Devadas

# Questions & Answers