DATA SCIENCE HEALTHCARE PROJECT

INDIVIDUAL PROJECT NAME: Health and Care

Name : Abina Azees

MAIL ID: abinaazees@gmail.com

Country : UAE

College: Amal Jyothi College of Engineering and Technology

Specialisation: Data Science

Problem Description

Pharmaceutical companies often struggle to assess how consistently patients adhere to their prescribed treatments, a concept known as **drug persistency**. This persistency plays a vital role in determining not only the **effectiveness of medical therapies** but also their **impact on patient outcomes** and the **commercial success** of the drugs.

To address this challenge, **ABC Pharma** has collaborated with an analytics firm to build a **predictive model** capable of identifying patients who are likely to be **persistent or non-persistent** with their therapies. The goal is to leverage a combination of **demographic, clinical, treatment-related, and adherence data** to uncover key factors influencing persistency. These insights will help enhance **patient engagement strategies**, optimize **treatment planning**, and ultimately improve **healthcare outcomes and business performance**.

Business Understanding

Drug persistency is a key indicator of how effectively patients adhere to their prescribed therapies. Gaining insights into persistency patterns provides valuable opportunities to enhance both **clinical outcomes** and **business performance**.

Why Drug Persistency Matters:

- **Improved Patient Targeting:** Enables more tailored and proactive interventions for patients at risk of non-adherence.
- **Optimized Sales and Marketing:** Facilitates data-driven strategies to promote therapies to the right patient segments.
- **Better Treatment Outcomes:** Helps reduce relapses and health complications by ensuring continuous medication intake.
- **Operational Efficiency:** Supports cost-effective resource allocation by focusing on patients who are less likely to persist.

Stakeholder Objectives:

- **ABC Pharma:**
    - Increase overall drug persistency rates
    - Enhance therapeutic effectiveness and patient satisfaction
    - Maximize market share and revenue growth
- **Analytics Partner:**
    - Build a robust, interpretable, and high-performing predictive model
    - Provide actionable insights into the drivers of persistency
    - Deliver a solution that can be seamlessly integrated into ABC Pharma's decision-making processes

Project Lifecycle

The project duration varies from June 19 to July 30.

Phase	Task Description
1 Problem Understanding	Define objectives, identify business and ML problems.
2 Data Understanding	Explore the dataset, identify features, understand variable types, distributions, and missing values.

Phase	Task Description
3 Data Cleaning & Feature Engineering	Handle missing values, encode categorical variables, normalize/scale numerical features, and create derived features if needed.
4 Model Development	Split data, train multiple classification models (e.g., Logistic Regression, Random Forest, XGBoost).
5 Model Evaluation & Selection	Evaluate using Accuracy, Precision, Recall, ROC-AUC. Choose best-performing model.
6 Model Deployment	Deploy the model using Flask or Streamlit (web app or API).
7 Reporting	Document process, create presentation, and write final report including: insights, performance metrics, feature importance, business impact, and challenges.
Final Submission	Submit GitHub repo, PDF report, and deployment link.

Data Intake Report

Name : Persistency of a Drug – Classification Model

Client: ABC Pharma

Date: June 19

Dataset Overview

Number of Observations: 3424

Total Features: 68

File Format: xlsx

Feature Group	Description	Example Variables
Patient Identification	Unique ID of patients	Patient_ID
Target Variable	Indicates persistency of therapy	Persistency_Flag
Demographics	Patient characteristics	Age, Race, Gender, Ethnicity, Region

Feature Group	Description	Example Variables
Provider Attributes	Prescriber-related features	NTM - Physician Specialty, IDN Indicator
Clinical Factors	Risk indicators and scan history	NTM - Risk Segment, Change in T Score, Dexa Scan Recency, Fragility Fracture During Therapy
Disease/Treatment Factors	Conditions and drug usage before/during therapy	NTM - Comorbidity, Glucocorticoid Usage, Injectable Experience, Concomitancy
Adherence	Overall adherence level	Adherence

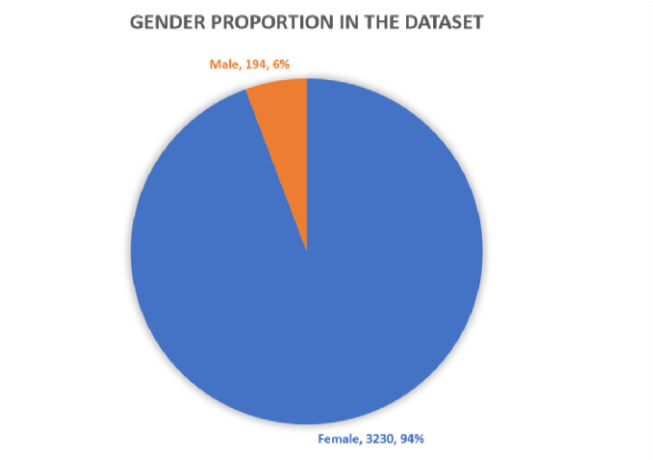
Proposed Approach

- Perfom EDA on the Dataset
- Handle missing values
- Encode categorical variables
- Normalize/scale numerical features (if required)
- Feature engineering for derived insights (e.g. patient risk profiles)
- Train/test split and model development

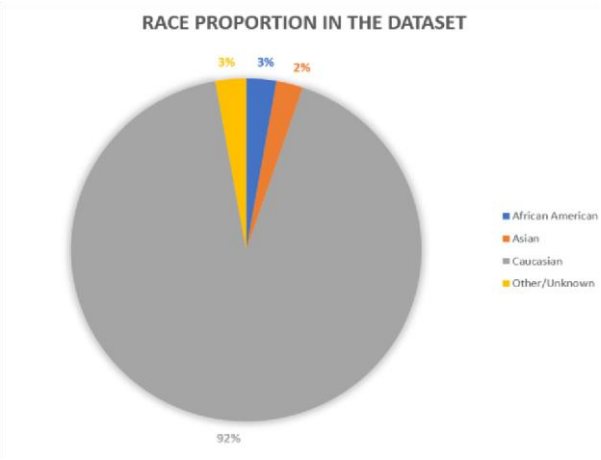
DEMOGRAPHICS

Demographic data including Gender, Race, Age, Region, Ethnicity, and IDN Indicator were analyzed in Microsoft Excel using Pivot Tables and visualized with Pivot Charts, as illustrated in Figure 2. The descriptive analysis reveals a notable demographic imbalance within the dataset. For instance, approximately 94% of the subjects are female, while only 6% are male, indicating a significant gender skew toward female patients.

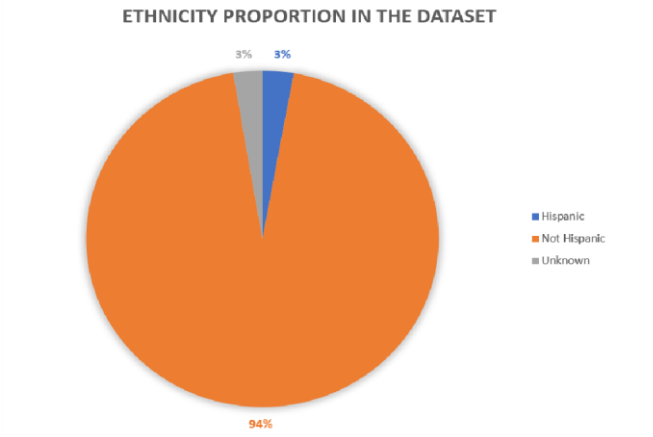
This imbalance raises an important question: is the dataset focused on a medical condition more prevalent among women, or does it reflect a trend where females—particularly within a specific age group—are more likely to seek medical care or adhere to treatment? Regardless, the dataset's gender bias suggests that any models or insights derived from it will be more representative of female patients and may not generalize well to the male population.



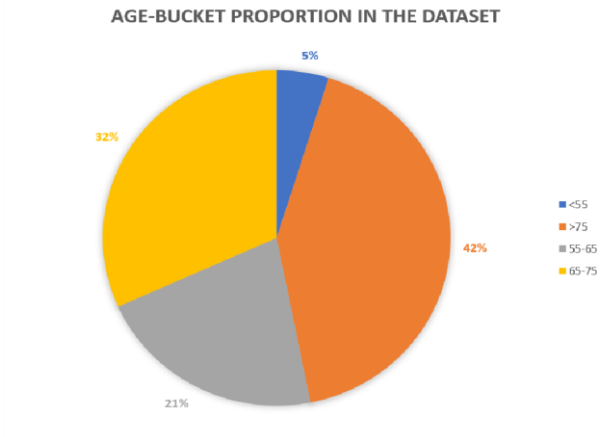
(a)



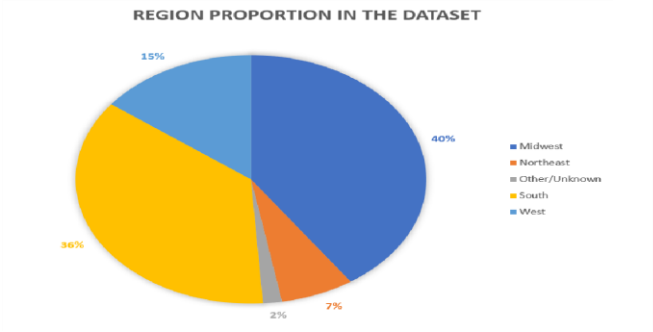
(b)



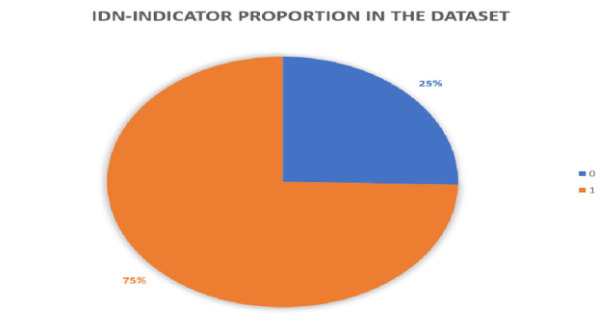
©



(d)



(e)



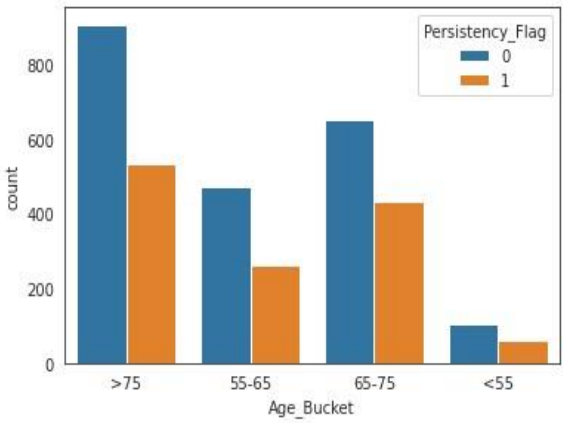
(f)

Figure 1: Visualization of the Descriptive Analysis on Demographics of the Subjects

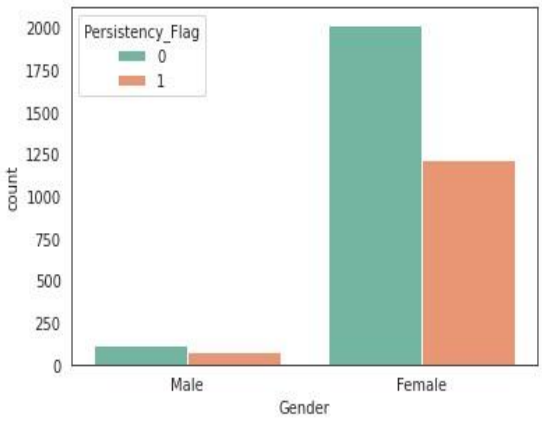
DEMOGRAPHIC PATTERNS

The dataset also shows a strong predominance of non-Hispanic, Caucasian patients, highlighting a potential lack of ethnic diversity. While the distributions across Region and Age categories were relatively more balanced, a closer look reveals that the majority of patients were over the age of 55, with very few younger individuals represented. Additionally, a large proportion of patients were from the Midwest region, indicating a geographic concentration in the data.

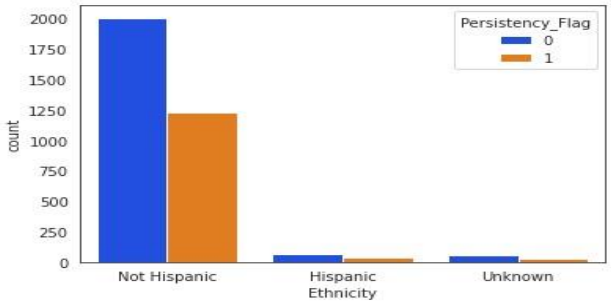
One of the most clinically significant findings is the high proportion of patients associated with an Integrated Delivery Network (IDN). Approximately 75% of the subjects were linked to a specific IDN, which may reflect the effectiveness of coordinated healthcare systems in improving patient access and experience. This high representation suggests the potential influence of organized healthcare networks on treatment patterns and outcomes.



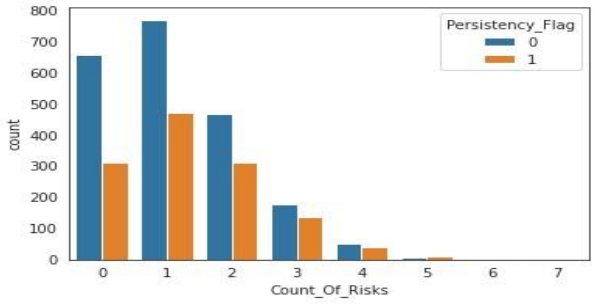
a



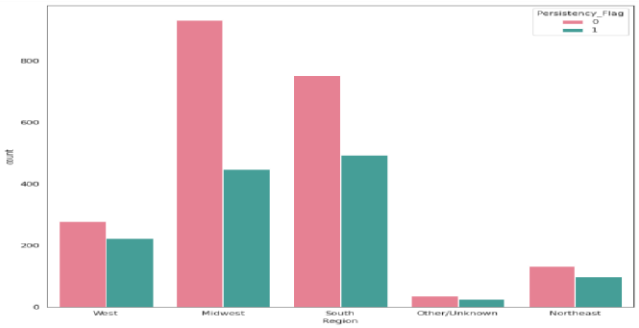
b



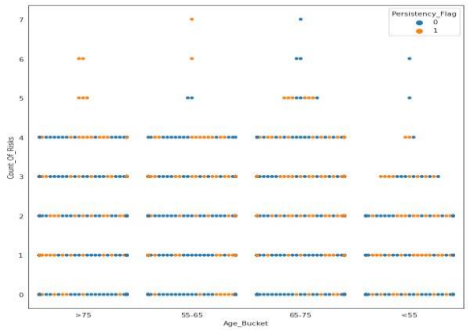
C



d



e



f

Figure 2: Visualization of Persistency Flags Related with Respect to Other Parameters

Practitioner Involvement in NTM Cases

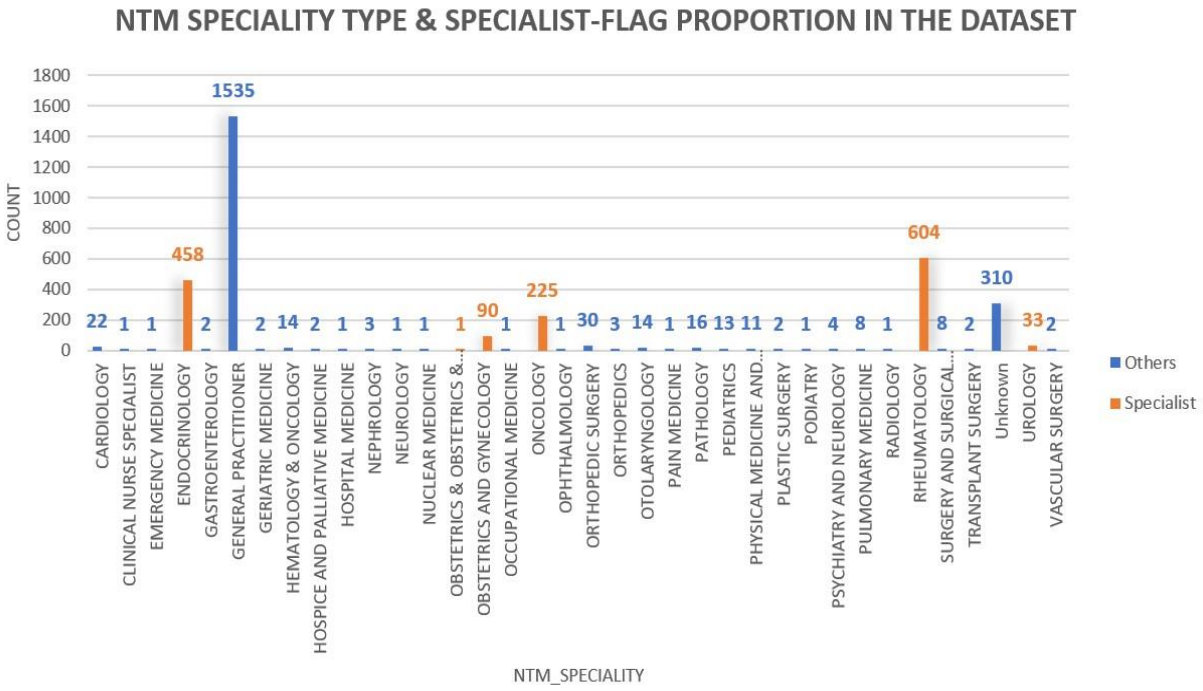


Figure 2: Specialty and Specialist-Flag of the Observing Medical Staff or Physician for the Patient

The analysis focuses on the healthcare practitioners involved in the cases, particularly examining their specialty types and whether they were classified as specialists in their respective fields.

A considerable portion of the physicians treating NTM cases were general practitioners, accounting for approximately 45% of all cases. Among the remaining practitioners, not all were identified as specialists. Based on the specialist flag indicator, only physicians in the fields of

Endocrinology, Obstetrics and Gynecology, Rheumatology, and Urology were recognized as specialists in their respective domains.

Data Transformation Approaches

The original dataset consists of both categorical (object-type) and numerical variables, which presents challenges for data analysis and modeling. In addition, some independent variables may be interrelated, leading to multicollinearity. From a linear algebra perspective, such redundancy can result in singular matrices, reducing the variability of the dataset and ultimately degrading the performance of predictive models.

To address these issues, several data transformation techniques were applied, including:

- Encoding categorical variables into numerical form using dummy variables,
- Removing highly correlated predictors to reduce multicollinearity,
- Generating lower-dimensional feature representations of the dataset.

Dummy Dataset Creation

As previously noted, out of 68 original features, only 2 were numerical while the remaining were categorical. To make the data suitable for machine learning algorithms, we applied one-hot encoding using pandas.get_dummies(). This process transformed each categorical variable into multiple binary (0/1) columns, resulting in a total of 119 features in the dummy dataset.

This transformation increased the number of total data points from **68 × 3424 = 219,136** in the original dataset to **119 × 3424 = 407,456** in the dummy dataset — nearly doubling the size.

The final dummy dataset, now fully numeric, was used in the machine learning pipeline for regression analysis. A total of 118 independent variables were utilized to predict the target variable, **“Persistency Flag”**.

This distribution suggests that NTM cases may be of greater clinical significance or complexity within these specialties. As such, their management may require specialized expertise to ensure accurate diagnosis and appropriate treatment pathways.

Turn your Categorical Column (Ex: "Name")...

Index	Name	8/6/2020
0	Liho Liho	$234.54
1	Chambers	$45.74
2	The Square	$56.22
3	Liho Liho	$32.31

...Into Dummy Indicator Columns

Index	Liho Liho	Chambers	The Square	8/6/2020
0	1	0	0	$234.54
1	0	1	0	$45.74
2	0	0	1	$56.22
3	1	0	0	$32.31

Figure 3: Encoding Techniques to be Applied on the Dataset

Feature Extraction

Autoencoders are effective tools for extracting meaningful features from large datasets by learning a low-dimensional representation, often referred to as a feature map. This compressed representation captures the essential information while reducing redundancy. The optimal number of features to extract depends on the sparsity and variability of the data; datasets with high redundancy can be efficiently represented with fewer features without significant loss of information.

Additionally, features can be selected manually and ranked based on their relevance to the target variable(s), ensuring that the most important predictors are emphasized in subsequent machine learning or statistical analyses.

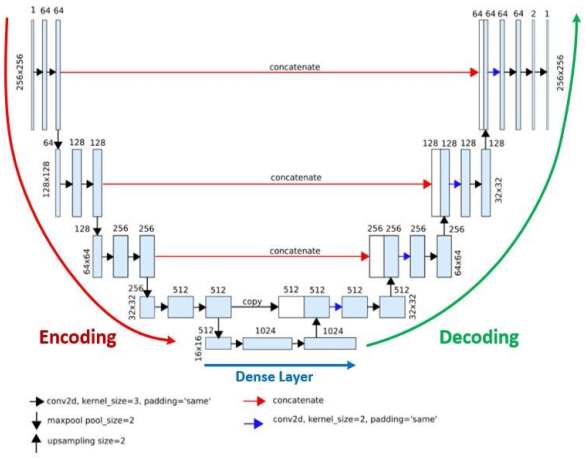


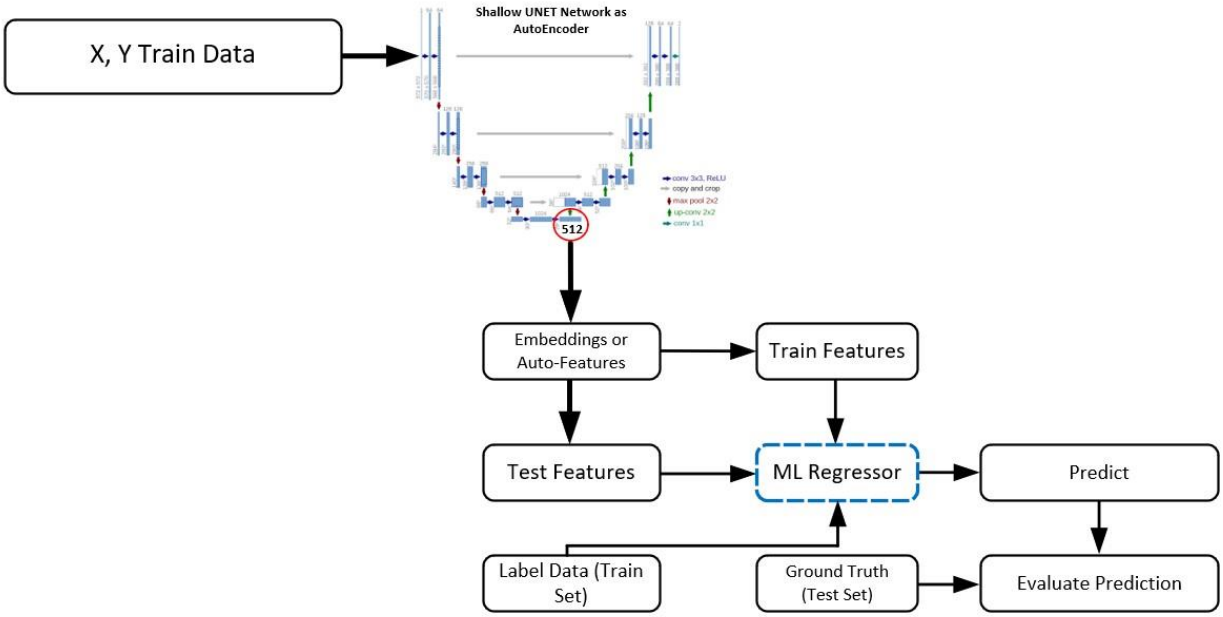
Figure 4: Structure for the U-Net based Autoencoder

Autoencoders perform feature extraction **automatically**, which is the origin of their name. There are several types of autoencoders, with the most common being Vanilla Autoencoders. For this project, however, we utilize the encoder portion of a U-Net deep segmentation network to encode the dataset into a compact, representative feature set. These features are extracted from the network's deepest dense layer.

Originally designed for image segmentation, the U-Net architecture has also been successfully applied to one-dimensional data tasks, such as reconstructing photoplethysmogram (PPG) signals into arterial blood pressure (ABP) waveforms . Figure 3 annotates the U-Net structure, highlighting its two main components: the encoder, which compresses the input into a low-dimensional feature set, and the decoder, which reconstructs the target output (signal or image) from these features. An optional dense layer is added to facilitate feature extraction from the autoencoder.

Determining the optimal number of features requires experimentation, as it depends on the dataset's variability—a property unique to each dataset. It is important to note that if the number of features needed exceeds the number of original independent variables, using an autoencoder may not be beneficial unless it leads to a significant improvement in performance. For instance, if the optimal feature count is 128 while the original dataset contains only 118 independent variables, the autoencoder's advantage is questionable.

A successful autoencoder experiment is one where the model maintains predictive performance while representing the dataset with a more compact feature set. The full pipeline for this feature extraction process.


X, Y Train Data
Shallow UNET Network as AutoEncoder
512
Embeddings or Auto-Features
Train Features
Test Features
ML Regressor
Predict
Label Data (Train Set)
Ground Truth (Test Set)
Evaluate Prediction