

LangTriage

Classification of Patient Free-Text Symptom Descriptions by Urgency

Interim Presentation

Team members:

Nofar Kedmi

Diana Akoshvili



Project Description

CHALLENGE:

Hospitals, especially emergency rooms, are overwhelmed with patients presenting a wide variety of medical complaints, often using informal language and inconsistent terminology.

Triage decisions must be made quickly to prioritize care and ensure that urgent cases are treated first — but this manual process is time-consuming and can lead to delays.

OUR SOLUTION:

Developing an NLP model based on a Large Language Model (LLM) to classify patient symptom descriptions into:

- 0 — **Not urgent:** Can wait or be monitored
- 1 — **Urgent:** Requires immediate medical attention

GOAL:

Support faster and more consistent triage decisions.

From Text to Triage Decision

Task:

- **Input:** Free text generated from structured clinical measurements.
- **Output:** A numerical label indicating urgency level (0 or 1).
- **NLP Tasks:** Text classification – Assign urgency labels to medical texts,
Data to Text Generation – from structured clinical data to symptom descriptions.

Data and evaluation:

- **Dataset:** Kaggle's Patient Priority Classification [dataset](#)
- **Labels** - Provided in the dataset as one of four categories:
 - Red: Requires immediate life-saving intervention
 - Orange: Serious symptoms, needs prompt medical attention
 - Yellow: Moderate symptoms, can wait briefly
 - Green: Mild condition, safe to wait
- **Evaluation:** Confusion Matrix and metrics: Accuracy, Precision, Recall, F1-Score, AUROC

Prior Art

Source/Title	Machine learning–based triage to identify low-severity patients with a short discharge length of stay in emergency department, 2022	Development of an Anticipatory Triage-Ranking Algorithm Using Dynamic Simulation of Patients With Trauma	Improving ED Emergency Severity Index Acuity Assignment Using Machine Learning and Clinical Natural Language Processing, 2021
Approach/Model	CatBoost, Logistic Regression, Random Forest	LIFE Priority algorithm	KATE triage model
Data	Triage records collected from two hospitals in Taiwan	An artificial database that included 82,277 patients with trauma injuries.	Electronic medical records from two US hospitals
Metrics	AUC, Precision	Comparison with START algorithm	Accuracy
Results	AUC = 0.755, Precision = 88.7%	LIFE improves the identification of clinical urgency compared to existing algorithms.	75.7% Accuracy

Pipeline

Mapping to Binary Labels:

Input: Raw urgency levels
('red', 'orange', 'yellow', 'green')

Process: Urgency levels were mapped to binary labels for classification:
red & orange → 1 (urgent)
yellow & green → 0 (not urgent)

Output: Binary label column for model training.



Generating a synthetic free-text complaint:

Input: Structured clinical indicators
(e.g. age, gender, blood pressure, glucose, chest pain type)

Process: Transformed into textual descriptions using GPT-4 model.

Output: Synthetic complaints stored in a new column called text_input, used as input to the language models.

.



Tokenization & vectorization:

Input: patient complaint.

Process:
For TF-IDF model: based on term frequency and inverse document frequency.
For DistilBERT model: Built-in tokenization (WordPiece) + contextual embedding vectors.

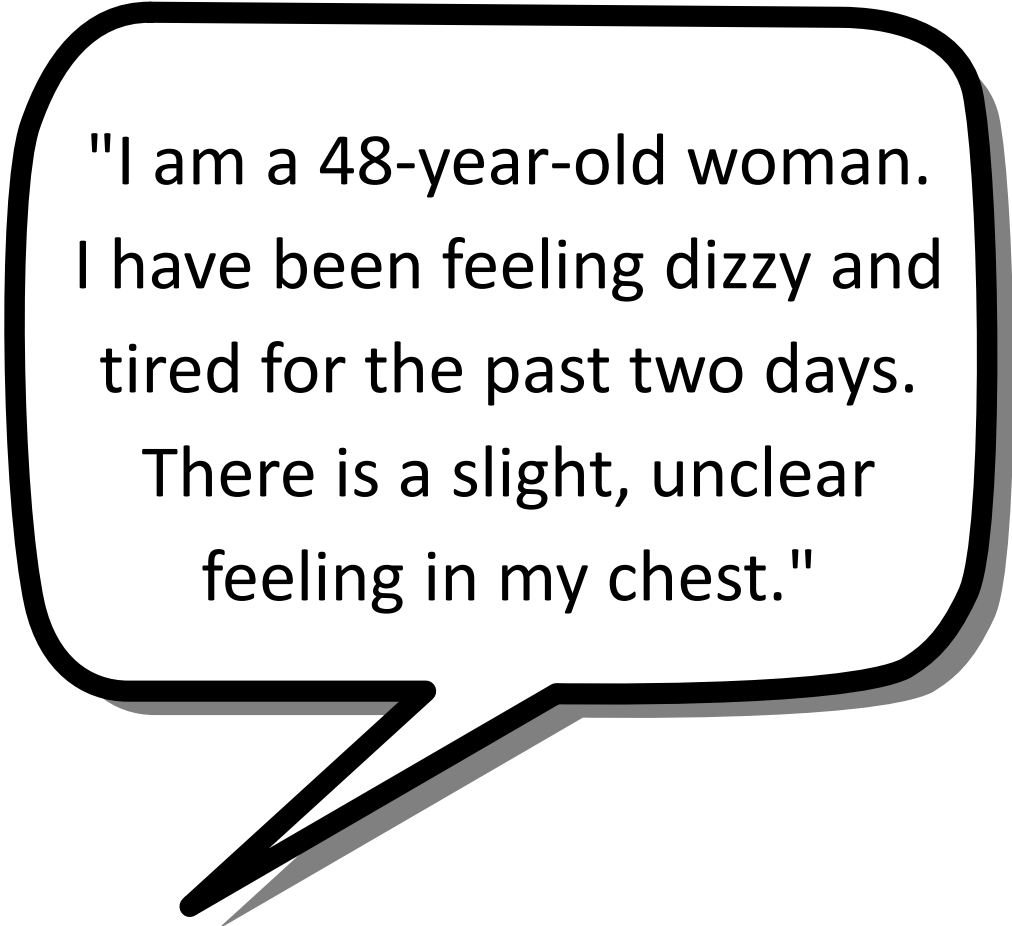
Output: Vectorized representation of each text sample for model input.

Structured Features & Patient Descriptions Overview


Main Features:

- Age
- Gender
- BMI
- Insulin
- Blood pressure
- Chest pain type
- Plasma glucose
- Exercise angina
- Heart disease
- Cholesterol
- Residence type
- Smoking status

Case Examples



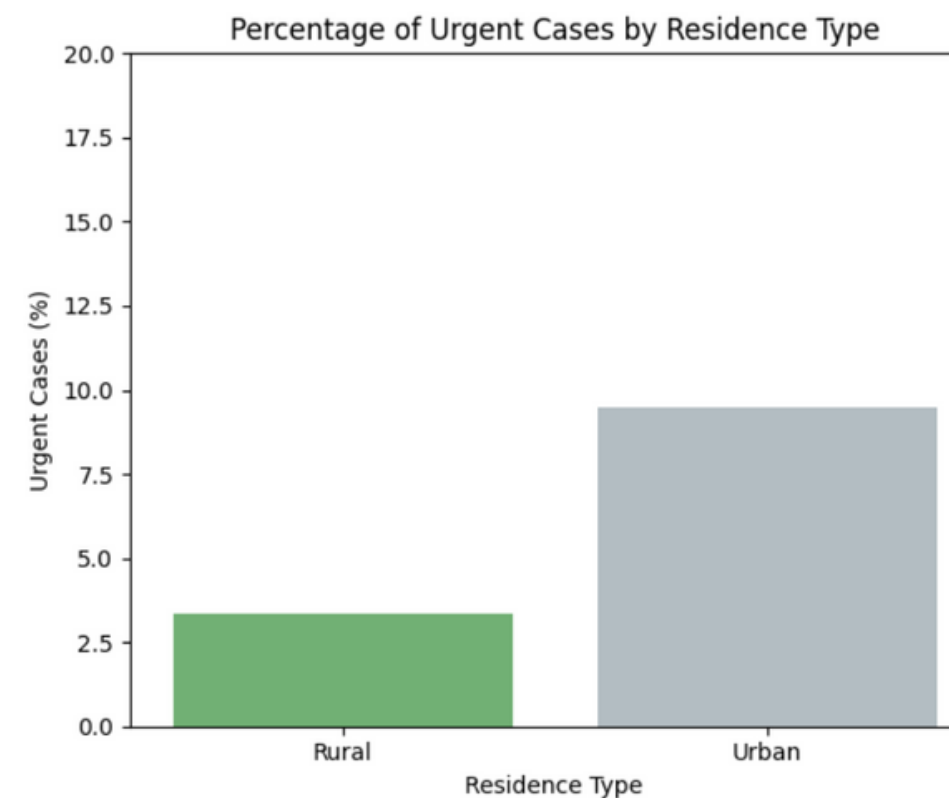
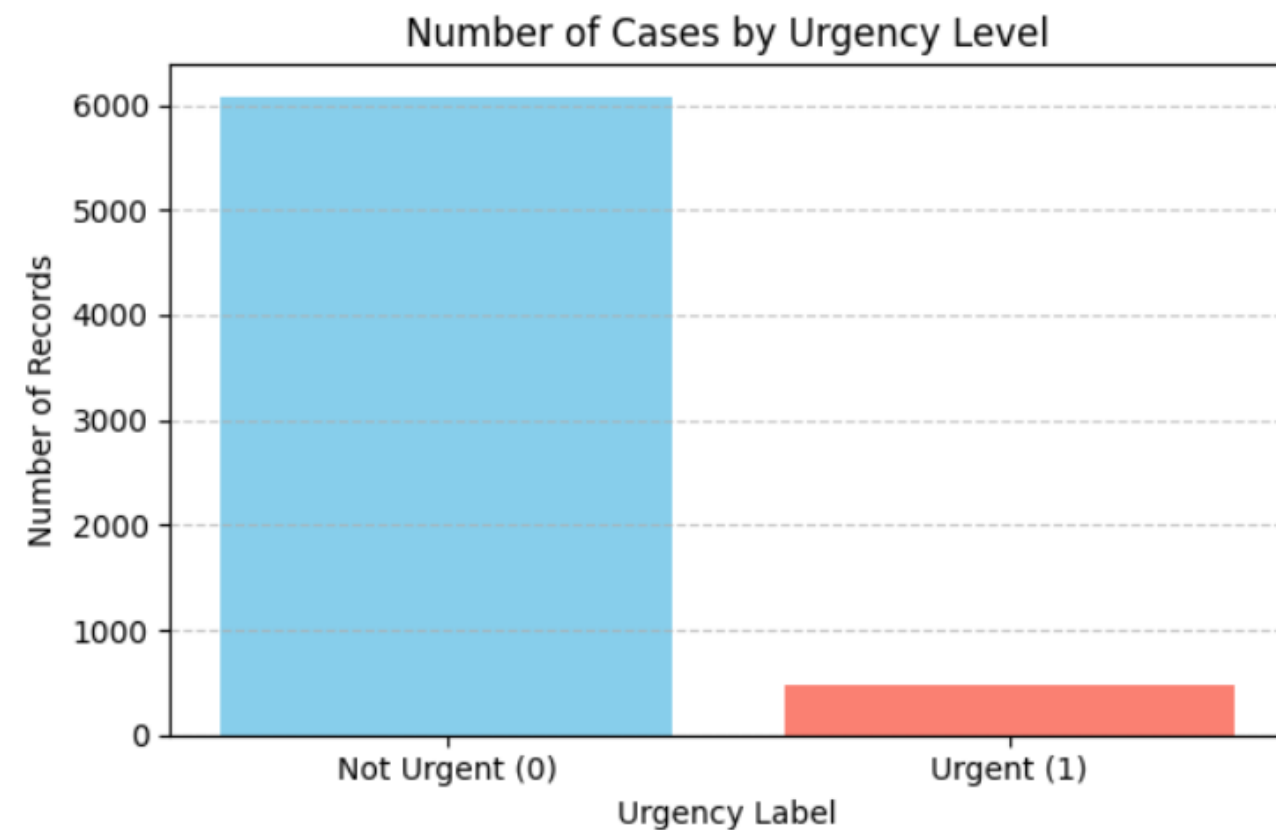
"I am a 48-year-old woman.
I have been feeling dizzy and
tired for the past two days.
There is a slight, unclear
feeling in my chest."



"I'm a 37-year-old man.
I feel blurred vision and
cannot concentrate.
I also feel very thirsty and go
to the bathroom more
frequently than usual"

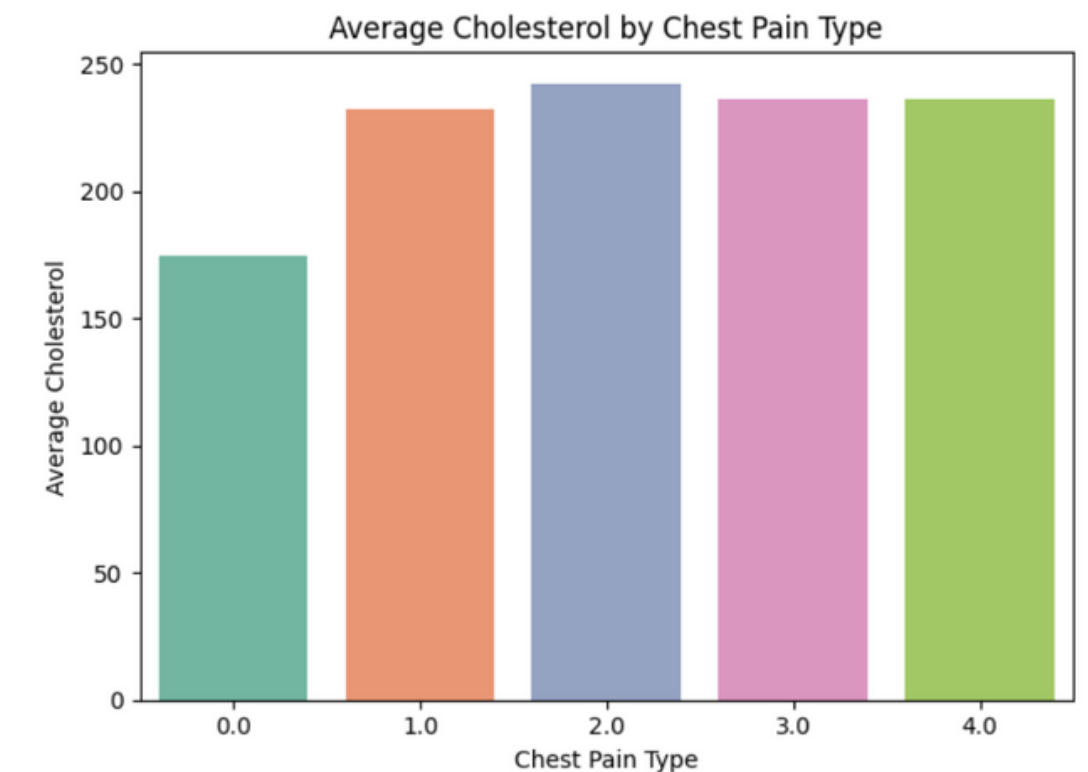
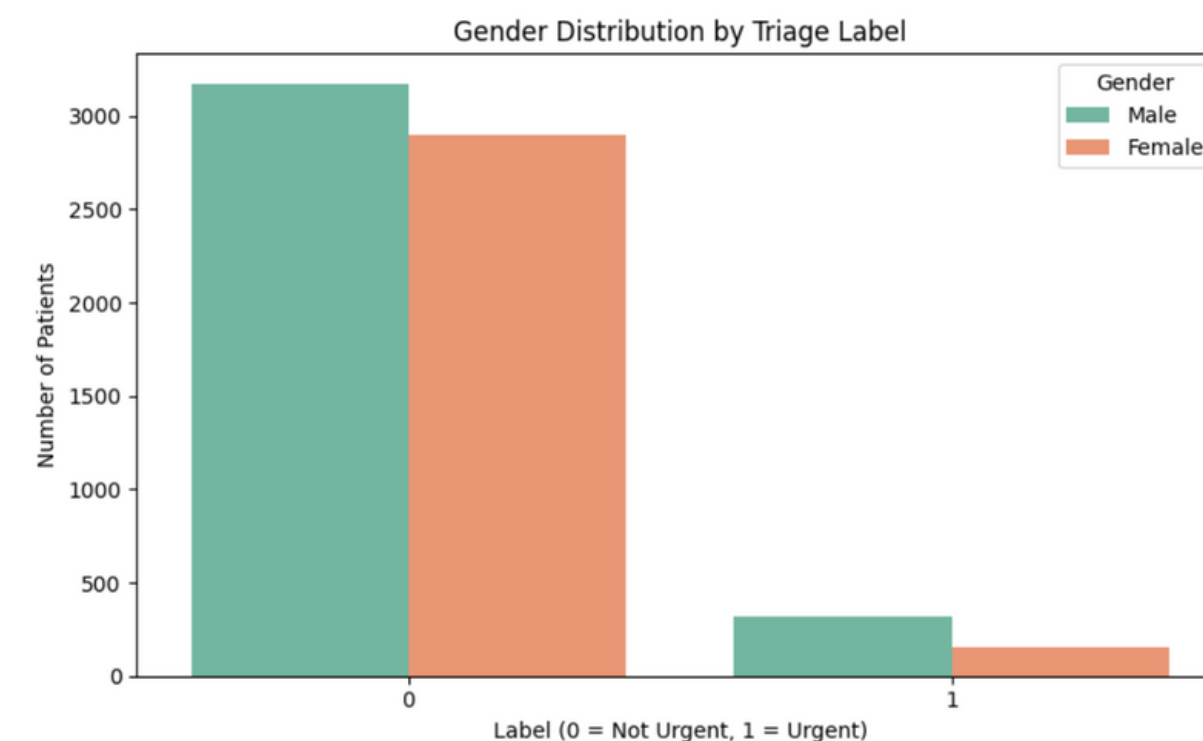
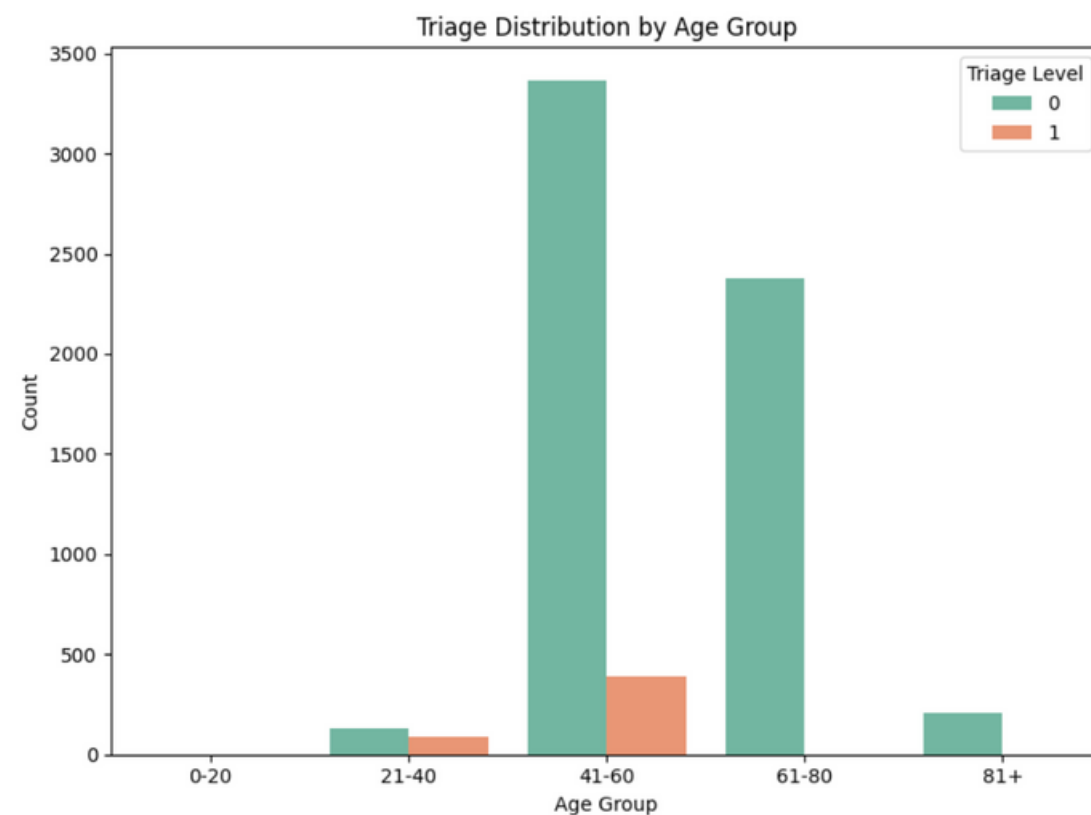
EDA

- **Dataset size:** 6962 patient records with 17 structured clinical and features.
- **Duplicate row check:** None in dataset.
- **Missing values:** Handled or removed to ensure clean input.
- **Class distribution:**
Urgent (1): 8% , Not urgent (0): 92%
- **Correlation:** presented using a correlation matrix. There are no significant relationships between variables and the target variable.



EDA - Insights

- **Middle age** (41-60) is likely to be the age with the highest risk for problems requiring immediate treatment.
- Among emergency patients (label = 1): **Men** constitute a larger group of emergency cases than women.
- There is a strong **association** between patients reporting **non-classical chest pain** (atypical angina) and high **cholesterol** levels - possibly due to a link to cardiovascular disease.
- A potential relationship was observed between **place of residence** and **urgency level**: the proportion of urgent cases was higher among patients living in urban areas compared to those in rural areas.



Models & Evalution

Models: Compare models:

Baseline Model – **TF-IDF + Logistic Regression:**

- A relatively simple model in which texts are converted to numeric vectors using TF-IDF (which measures the importance of words in a document) and then classified with logistic regression.

Advanced Model – **Fine-tuned DistilBERT:**

- Enhanced Transformer language model (DistilBERT) fine-tuned on our data. This is a model that understands language context at a high level.

Data splitting:

- Dividing into 80% training and 20% testing.

Evalution:

- Confusion Matrix and metrics: Accuracy, Precision, Recall, F1-Score, AUROC

Baseline - Overview

Text vectorization:

Converting patient complaints into numeric vectors, by calculating the frequency of words relative to other documents (TF-IDF). Both single words (unigrams) and word pairs (bigrams) are included.

Feature restriction:

The top 1,000 most informative words and word pairs are selected to reduce dimensionality (max_features=1000).

Model Training and Prediction:

A logistic regression classifier is trained on the vectors to predict the urgency level of each case (0 = Not Urgent, 1 = Urgent).

Baseline - Results

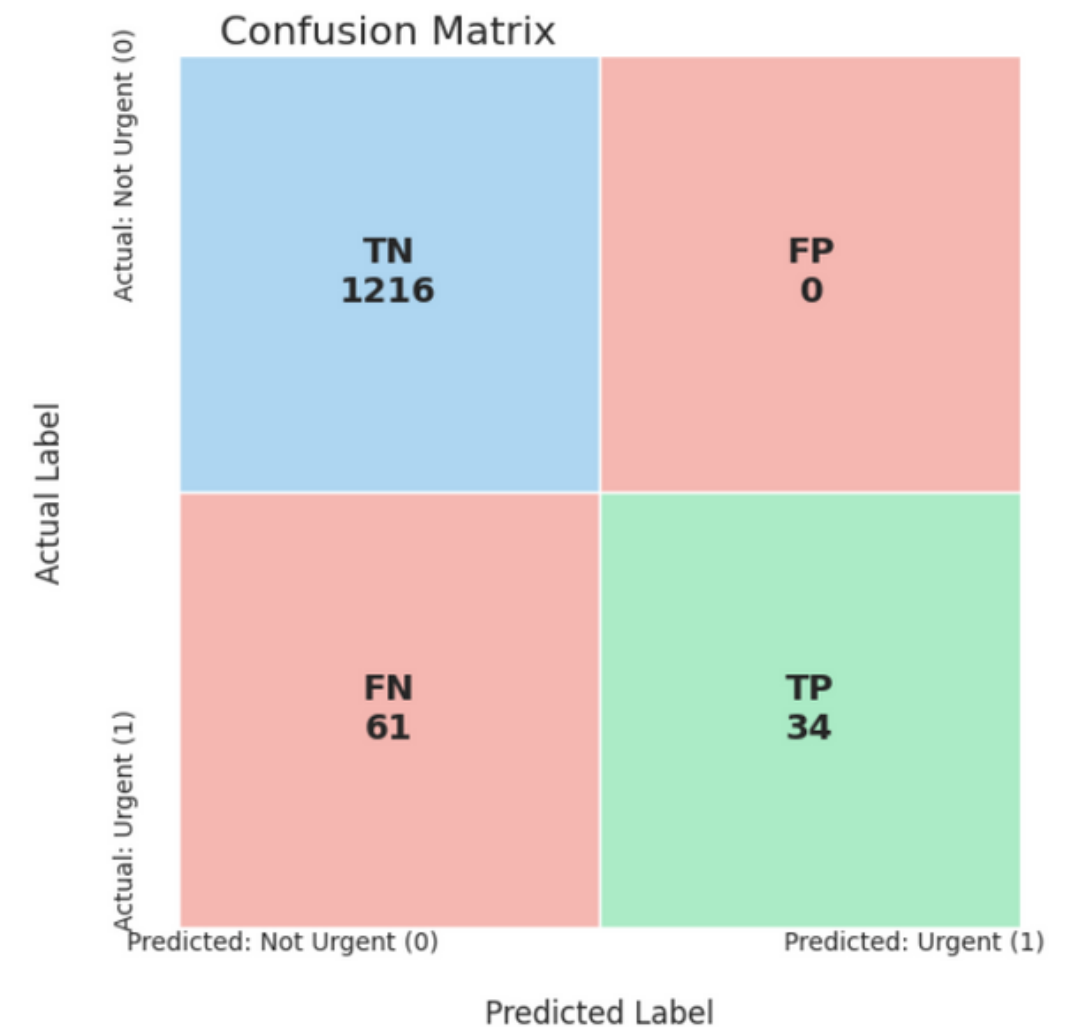
Accuracy (0.95): The model classifies most samples correctly.

Recall (0.36): Only 36% of urgent cases were detected — out of 95 urgent cases, 61 were missed.

Precision (1.00): When a case is predicted as urgent, it is almost always correct — preventing false alarms.

F1-Score (0.53): Overall performance is moderate due to low recall.

AUROC (0.91): Good discrimination between urgent and non-urgent cases.



Conclusions and Next Steps

The basic model successfully identifies non-urgent cases but misses many urgent ones, which can be dangerous in clinical settings. This issue is likely caused by an imbalance between urgent and non-urgent cases.

Future Recommendations:

1. Resampling the Categories – Apply techniques to address class imbalance:
 - SMOTE: Generate synthetic examples of urgent cases.
 - Undersampling: Reduce the number of non-urgent examples.
 - Combination of both: Achieve a better balance between the classes.
2. Model-Level Adjustments – Improve recall without altering the dataset:
 - Adjust the cut-off threshold or tuning class weights.