Virtual Care Assistant from home

Presentation Part three

Presented By :

Shay Sason
Gabrielle Maor

Problem Description

Monitoring Changes in the Condition of Home Hospitalization Patients Using Clinical Data and Free Text

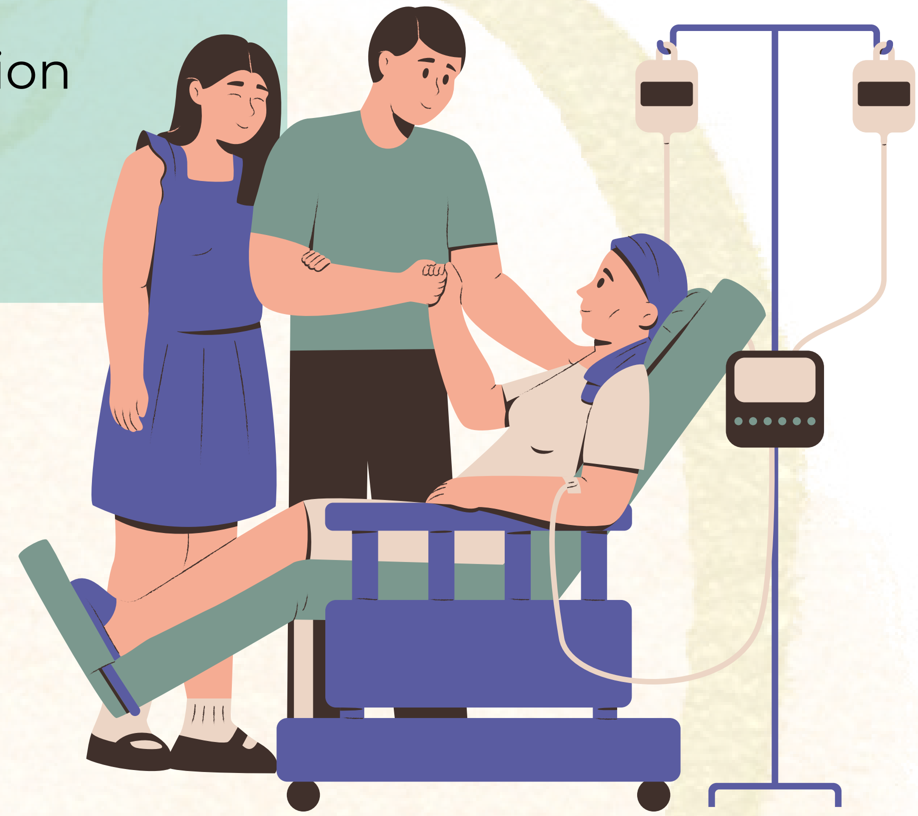
In the context of home hospitalization, medical staff monitor the patient's condition through daily visits and regular reports.
Data is collected from various sources – physiological measurements, medical documentation, and free-text descriptions from both patients and caregivers.

The challenge:

The information is dispersed, inconsistent, and often difficult to piece together into a clear clinical picture.
Changes in the patient's condition - whether improvement or deterioration - can go unnoticed.

Key Problem:

How can we automatically detect day-to-day changes in a home-hospitalized patient's medical condition by combining free-text input with structured clinical data?



Project Objectives

Effective home-patient monitoring demands accurate, timely clinical insights. This project aims to develop a clinical decision support tool for home hospitalization teams by:

- Predicting changes in a patient’s condition between visits (improvement, no change, or deterioration).
- Integrating heterogeneous data sources, including unstructured clinical notes and structured physiological measurements.
- Demonstrating the effectiveness of a hybrid NLP+ML model in automating medical monitoring and detecting significant health trends.

1	2	3	4	5	6	7	8
Data Exploration & Construction	Data Cleaning & Preprocessing	Exploratory Data Analysis (EDA)	Feature Engineering	Modeling	Evaluation	Visualization & Reporting	Conclusions & Future Work

Task Specification

Input:

- Free-text reports from home hospitalization on Day 1 and Day 2
- Clinical measurements: heart rate, respiratory rate, blood pressure, temperature, oxygen saturation.

Output:

- Classification of patient status: Deterioration, No Change, or Improvement

Performance Metrics:

- Accuracy
- F1-score

Plan:

1. Data Preparation & EDA
2. Text & Vitals Feature Engineering
3. Baseline Modeling (Text or Vitals only)
4. Combined Modeling (Text + Vitals)
5. BERT-based Models
6. Experimental Fusion Architectures
7. Final Evaluation & Model Selection

Project Plan

1. Data Preparation & EDA

- Synthetic dataset generation simulating home-hospitalization cases
- Manual labeling of condition change: Improvement / No Change / Deterioration
- Initial class balance analysis
- Split into Train / Validation / Test (80/20) with stratification

2. Text & Vitals Feature Engineering

- Merge day1 and day2 notes → combined_text
- Text cleaning: lowercasing, punctuation removal, spell correction, stopword removal
- Vitals: compute Δ (day2 - day1) for HR, BP, Temp, etc.
- Imputation (SimpleImputer) and Standardization (StandardScaler)

3. Baseline Modeling (Text or Vitals only)

- TF-IDF + Logistic Regression
- Vitals only with XGBoost and Random Forest
- Evaluate on validation/test splits using Accuracy & F1 (macro)

4. Combined Modeling (Text + Vitals)

- Fusion pipelines combining TF-IDF or BERT embeddings + ΔVitals
- Models: Logistic Regression, XGBoost, LightGBM, Neural Networks
- Cross-validated performance: Accuracy, F1 per class

5. BERT-based Models

- Manual fine-tuning on small dataset (BERT tokenizer + classifier head)
- Classifiers: Logistic, XGBoost, Neural Net
- Evaluate effect of pretrained embeddings vs. end-to-end finetuning

6. Experimental Fusion Architectures

- Concatenation of BERT pooled embeddings + vitals as input
- Neural MLP layers trained to classify 3 classes
- Improved architectures showed inconsistent results (precision/recall variance)

7. Final Evaluation & Model Selection

- Compare all models: Fusion, Classical, BERT-based
- Visualize results using Accuracy, Macro-F1, and per-class Recall
- Select best-performing model: LightGBM (TF-IDF + ΔVitals)
- Present Recall comparison across classes (improvement / deterioration / no change)

Prior Art

Source / Title (Year)	Approach / Model	Data (size)	Metrics	Results
"Deep learning for early warning of inpatient deterioration" (Nature Digit. Med., 2021) https://did.li/1otx5	LSTM with attention on vital-sign time-series	2,100 patient episodes (minute-by-minute vitals)	AUC, Macro-F1	AUC 0.87, Macro-F1 0.71; attention maps showed SpO_2 drops and HR spikes as strongest early-warning signals; model predicts deterioration up to 12 h in advance.
"Fusion of clinical notes and vitals using BERT + MLP" (J. Gen. Intern. Med., 2023) https://did.li/fAVaa	BERT encoder on free-text notes + MLP head for vitals	800 home-hospitalization episodes	Accuracy, Precision, Recall	Accuracy 0.76; Precision 0.74; Recall 0.72; text+vitals fusion improved Macro-F1 by 8 pp over vitals-only baseline; ablation confirmed synergy of both modalities.
"Combining vital signs and free-text for deterioration detection" (PMC, 2019) https://did.li/dR7CN	TF-IDF vectorization on clinical notes + LightGBM	1,200 inpatient stays (notes + vitals)	ROC-AUC, Sensitivity, Specificity	ROC-AUC 0.79; Sensitivity 0.81; Specificity 0.75; top predictive text tokens included "pain," "breathlessness," "fatigue"; combining features outperformed single-modality.

Data & Labeling

Synthetic Data and Clinical Status Labeling – Home Hospitalization

Data Generation

Patient Simulation:

The dataset was generated using gpt-4o prompting, simulating daily reports of home-hospitalized patients.

Key Fields:

- day1_note and day2_note: first-person, simple, everyday-style text
- Daily vitals: HR, BP, Temp, RR
- Status change direction – aligned with both free text and vitals
- reasoning: objective analysis of change between the two days

Randomized parameters:

Age (25-110), gender, medical diagnosis

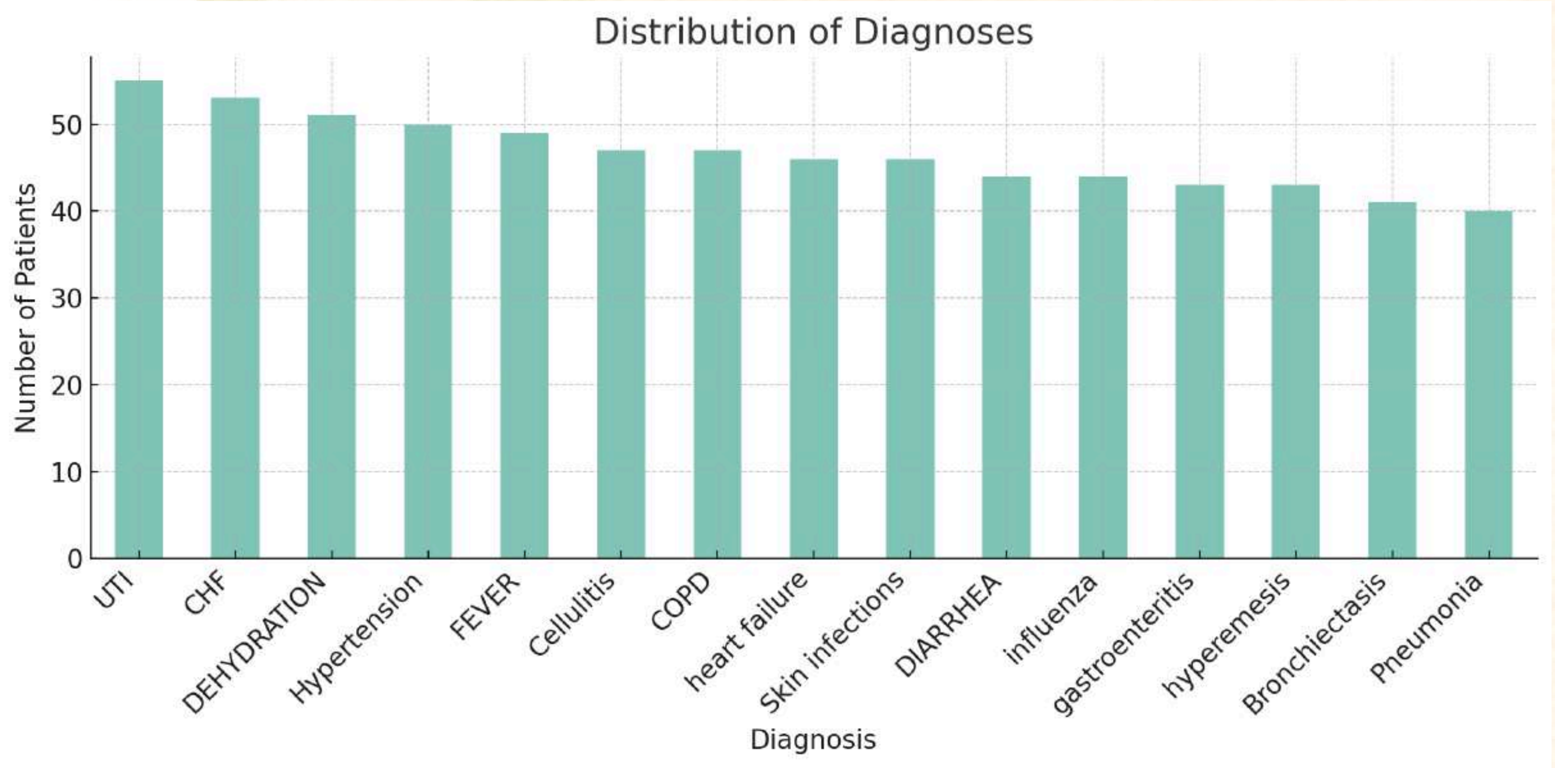
Labeling & Classification

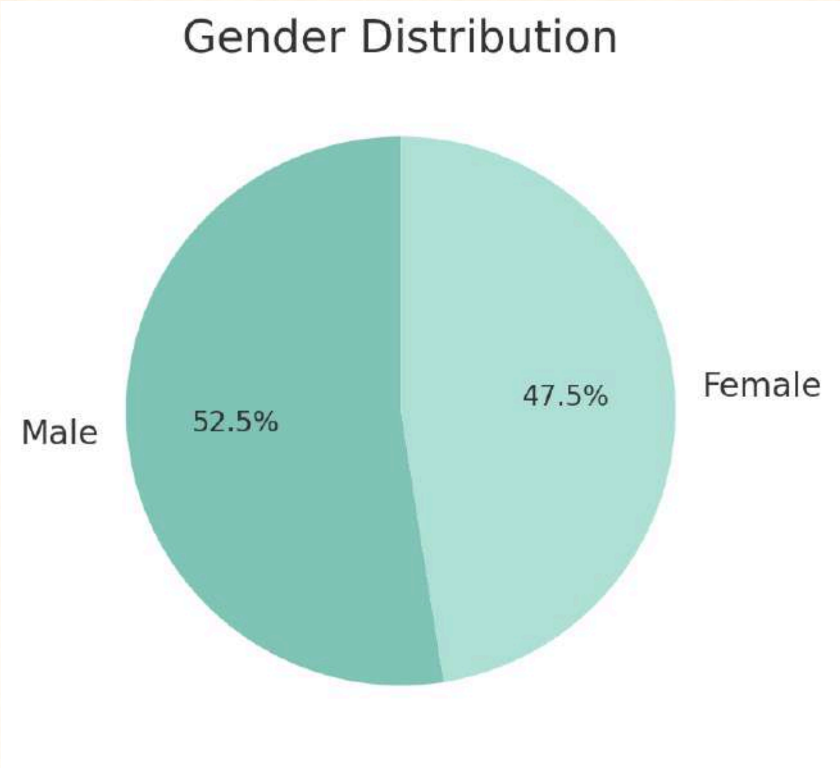
Classification Categories:

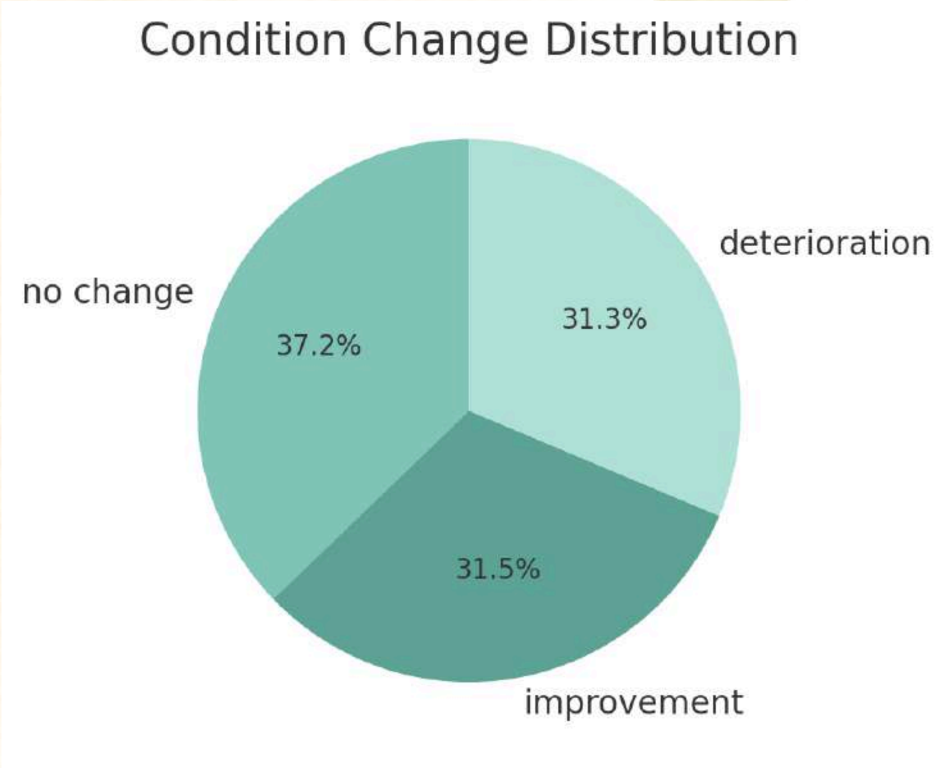
- Improvement
- Deterioration
- No Change

Labeling Criteria:

- Analysis of changes in vitals between Day 1 and Day 2
- Language cues in the notes indicating improvement or deterioration







Data Preprocessing

Preprocessing of Text and Physiological Data

Free Text (Complaints Notes):

- Merged day1_note and day2_note into a unified field: combined_text.

Applied multi-step text cleaning:

- Lowercasing, punctuation and digit removal.
- Spelling correction using SymSpell.
- Removal of stopwords via NLTK.
- Final result stored in cleaned_text.

NLP Techniques:

- TF-IDF vectorization (up to 500 features, bigrams included).
- Optional input to BERT tokenizer (bert-base-multilingual-cased) for deep models.
- Final text representations used for classical models or combined with vitals.

Physiological Data:

- Extracted vital signs from raw text (HR, RR, Temp, BP_SYS, BP_DIA) using regular expressions.
- Computed day-to-day differences (*_diff columns).
- Missing values imputed using SimpleImputer (mean strategy).
- Standardized via StandardScaler before modeling.

Outcome:

- Each patient is represented by a single feature vector, combining objective (vitals) and subjective (text) information.
- Enables robust multi-class classification into health status change (target = no change / improvement / deterioration).

Modeling & Pipeline

Classify changes in patient condition based on clinical notes and vital sign deltas

Processing Pipeline:

Text data was combined with physiological measurements using a scikit-learn pipeline:

- TF-IDF applied to combined_text
- SimpleImputer and StandardScaler used for vitals (HR_diff, RR_diff, Temp_diff, BP_SYS_diff, BP_DIA_diff)
- Spelling correction and stopword removal were applied to the text (for certain models)

Selected Models:	
Basic, easy to interpret, also used in the combined (fusion) model	**Logistic Regression**
Powerful model for tabular data; tested on vitals alone and on text+vitals	**XGBoost**
Fast alternative to XGBoost; also tested in the combined model	**LightGBM**
Text tokenized using bert-base-multilingual-cased for future integration	**BERT (Tokenization)**
TF-IDF on text + vitals → Logistic Regression	**Fusion Pipeline**

Key Notes:

- Data was split into Train/Validation/Test sets using Stratified Split to preserve label distribution
- 5-fold Cross Validation was used for performance evaluation
- A comparative analysis was conducted between raw text and cleaned/corrected text
- Models were kept as simple as possible to ensure clinical interpretability

Evaluation Metrics & Performance

Key Metrics:

Accuracy:

The proportion of correct predictions out of all samples.

A basic and convenient metric for overall comparison.

Macro F1-Score:

The harmonic mean of Precision and Recall, calculated separately for each class and then averaged.

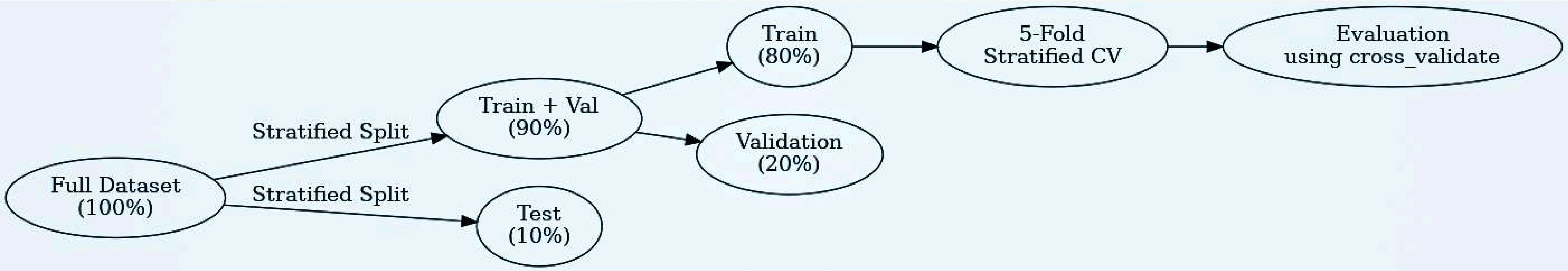
Particularly useful in cases of class imbalance.

Precision & Recall:

- Precision: The percentage of correct predictions out of all instances predicted for a given class
- Recall: The percentage of actual class instances correctly identified by the model

Evaluation Methodology:

- An early split was performed into Train/Validation/Test sets: 90% Train+Val / 10% Test, using stratified split
- Within the Train+Val set, an additional split of 80% Train / 20% Validation was applied
- For some models, 5-fold Stratified Cross Validation was used
- Metrics were computed using cross_validate from scikit-learn, without relying on classification_report



Intermediate & Baseline Results

Baseline Models (Train+Val, 5-fold CV):

Model	Input	Accuracy	F1 Macro
Logistic Regression	Raw Text (TF-IDF)	0.951	0.951
Logistic Regression	Cleaned Text	0.896	0.897

Conclusion: Even a simple text-only model performs well, but excessive text cleaning reduces performance.

Process:

Input:

Free-text from medical records

Text Processing:

- TF-IDF Vectorization – Extracting features based on word frequency

Two Experiments:

- Raw Text – No cleaning applied
- Cleaned Text – Stopword removal and spelling correction

Fusion Model – Text + Vitals Deltas:

Set	Accuracy	F1 Macro
Train+Val	0.954	0.954
Test	0.943	0.943

Conclusion: Significant improvement over text-only models. Integrating vital sign deltas adds clear value.

Process:

Input:

Clinical Text → Processed using TF-IDF

Physiological Measurements → Calculated as delta values (change between consecutive readings)

Preprocessing:

Imputer to handle missing values

StandardScaler to normalize the vitals

Fusion:

A unified pipeline combining both text features and vital sign deltas

Model:

Logistic Regression – Classic and interpretable

Main Results – Final Model Comparison

Model Comparison Insights

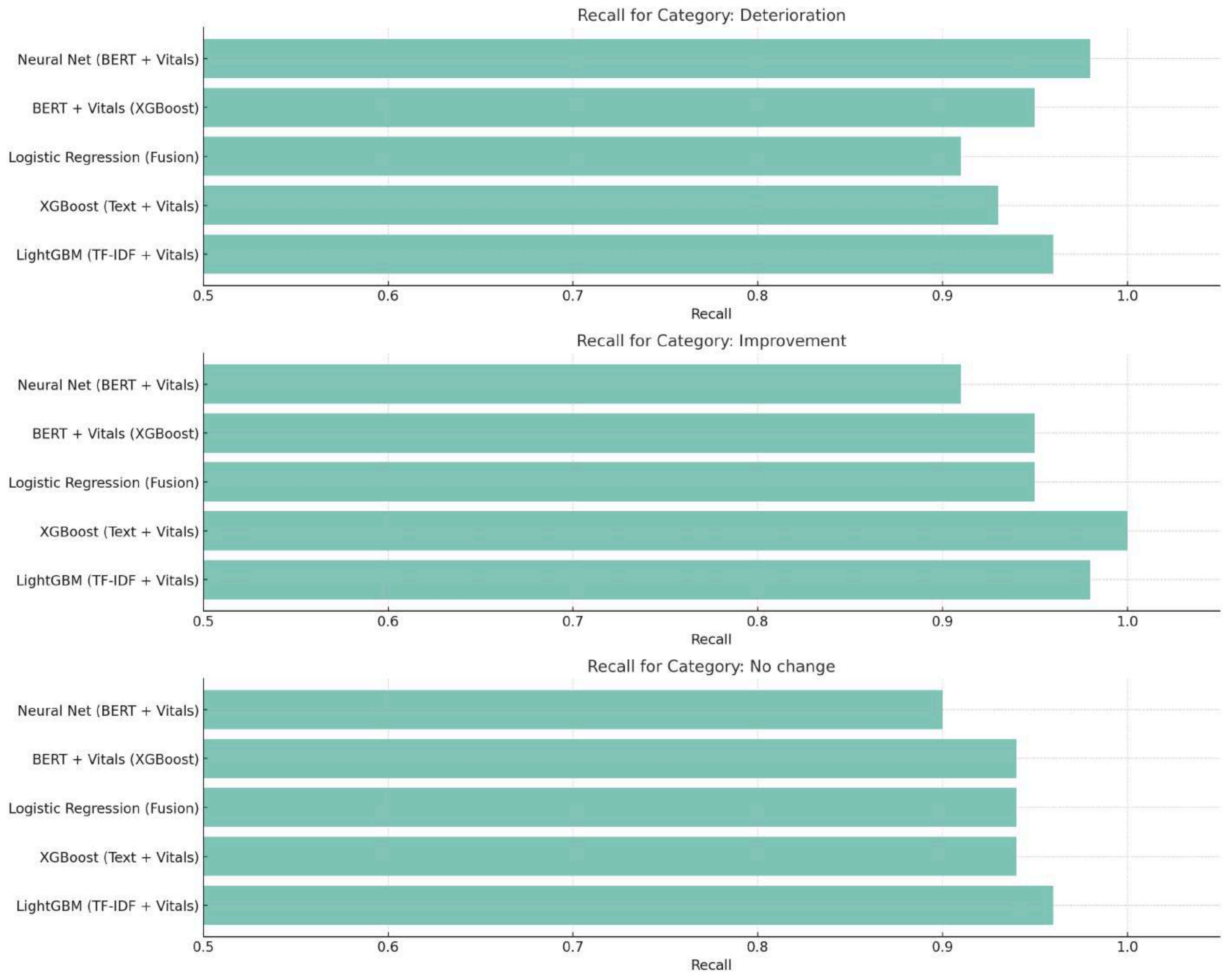
- Combined models (Text + Vitals) achieved the highest performance, especially with XGBoost and LightGBM.
- The Fusion Logistic Regression model delivered surprisingly strong results given its low complexity.
- BERT-only models did not consistently outperform others, but performed well when integrated with quantitative data (Vitals).
- Neural network models were less consistent — possibly due to sensitivity to hyperparameters or insufficient data.

Conclusion:

The best-performing models are those that combine natural language input with measured clinical data.

Choosing XGBoost or LightGBM with fused inputs offers the optimal balance between accuracy and interpretability.

Model	Accuracy	F1 Macro
LightGBM (TF-IDF + Vitals)	0.971	0.972
XGBoost (Text + Vitals)	0.971	0.972
Logistic Regression (Fusion)	0.943	0.943
BERT + Vitals (XGBoost)	0.943	0.944
Neural Net (BERT + Vitals)	0.936	0.937



Insights & Final Takeaways

1. Simple models can go a long way

Logistic Regression with TF-IDF alone achieved high performance (F1 = 0.95) on raw clinical text.

Takeaway: Even simple solutions can yield strong results, especially when the input text is rich in clinical context.

2. Data fusion is key to success

All models that combined text with vital signs showed a clear performance boost.

This fusion leverages both the narrative context and the patient's physiological state.

3. XGBoost and LightGBM stood out for both accuracy and consistency

These models reached over 97% accuracy when fed with combined inputs.

Their advantages include: robustness to missing data, feature importance analysis, and efficient learning.

4. BERT isn't always better — it's about how you use it

BERT alone performed lower (~81%), but when integrated with vitals, it achieved very strong results.

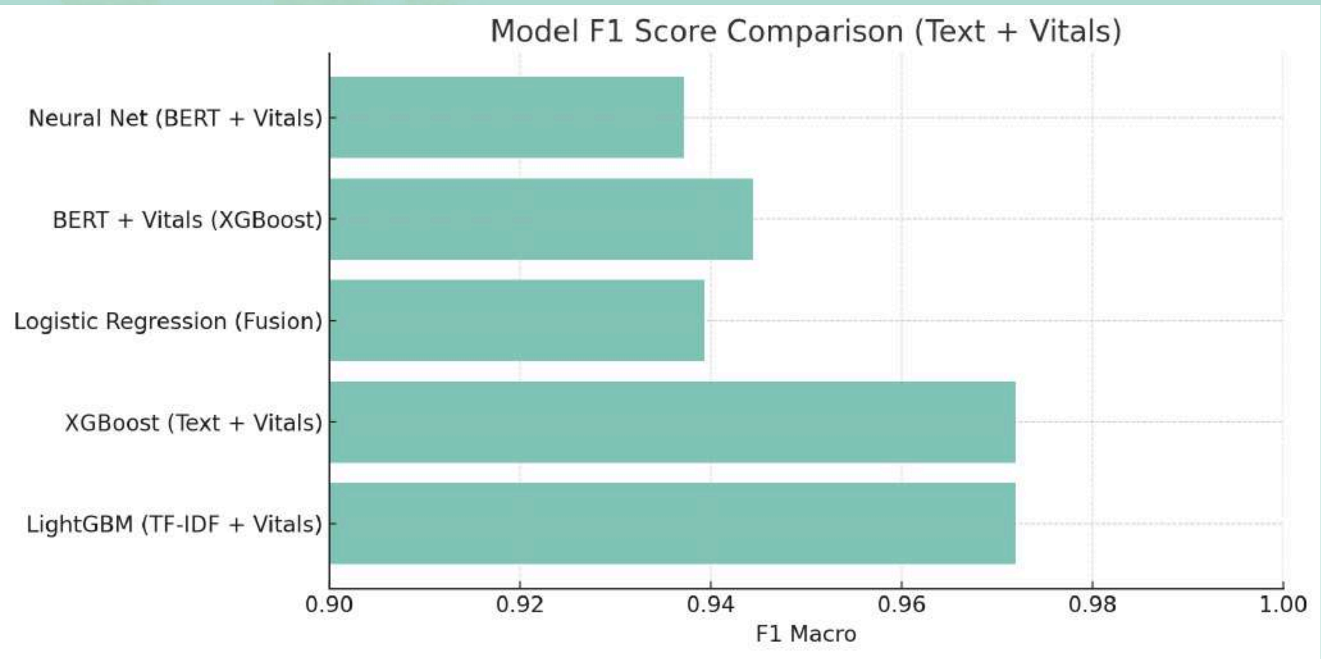
Conclusion: Smart integration matters more than the model's complexity.

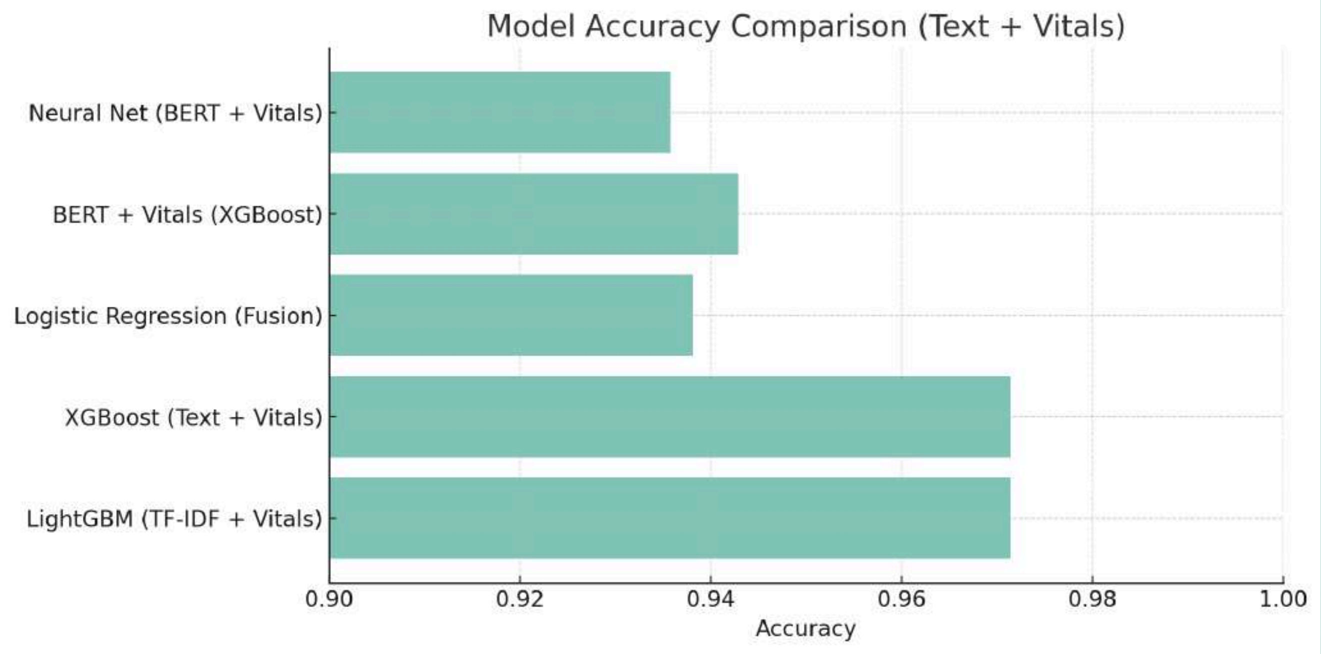
5. Deep models need optimization

Neural Networks showed potential but lacked consistency — likely due to data size limitations and sensitivity to hyperparameters.

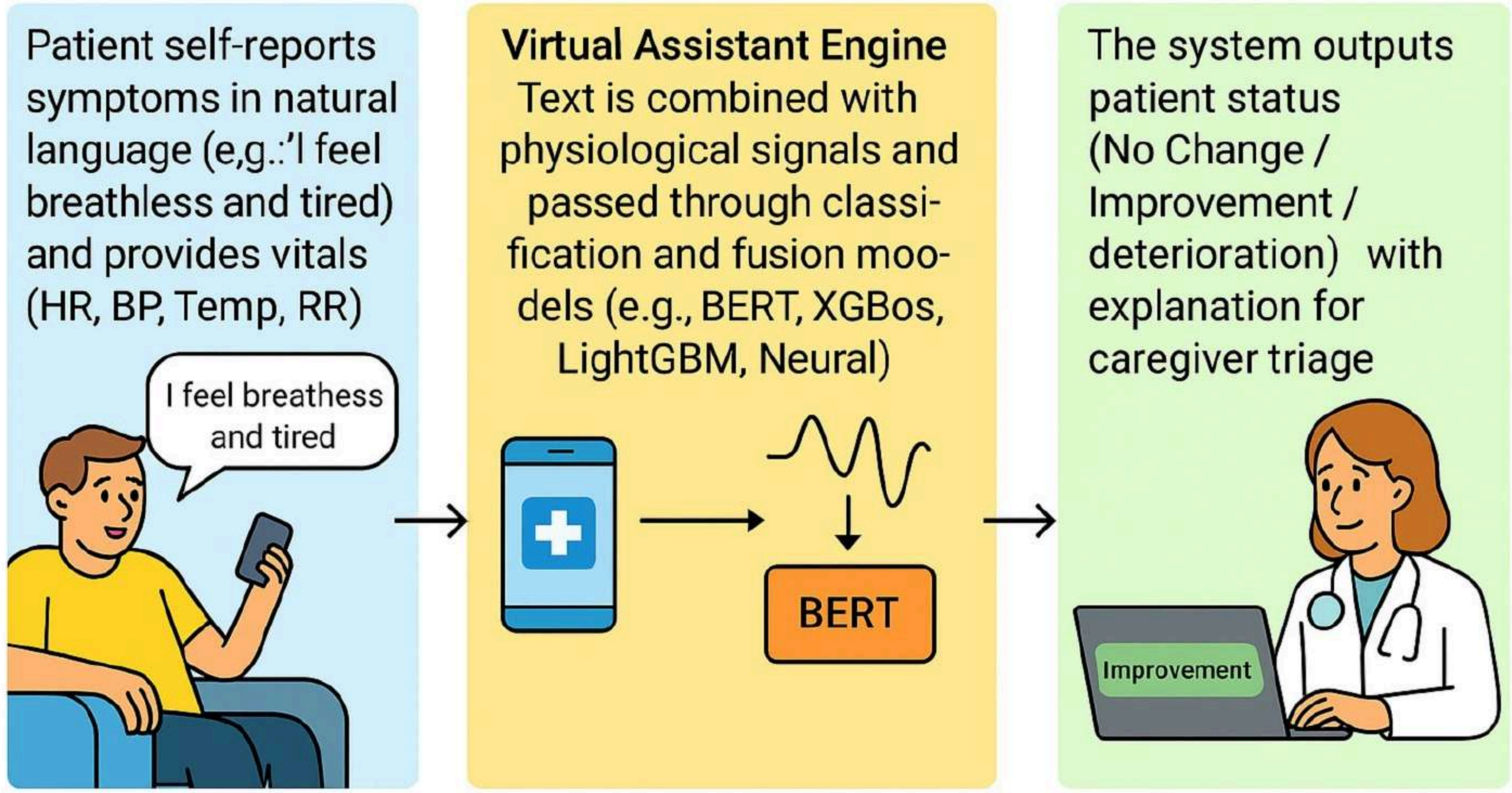
Conclusion:

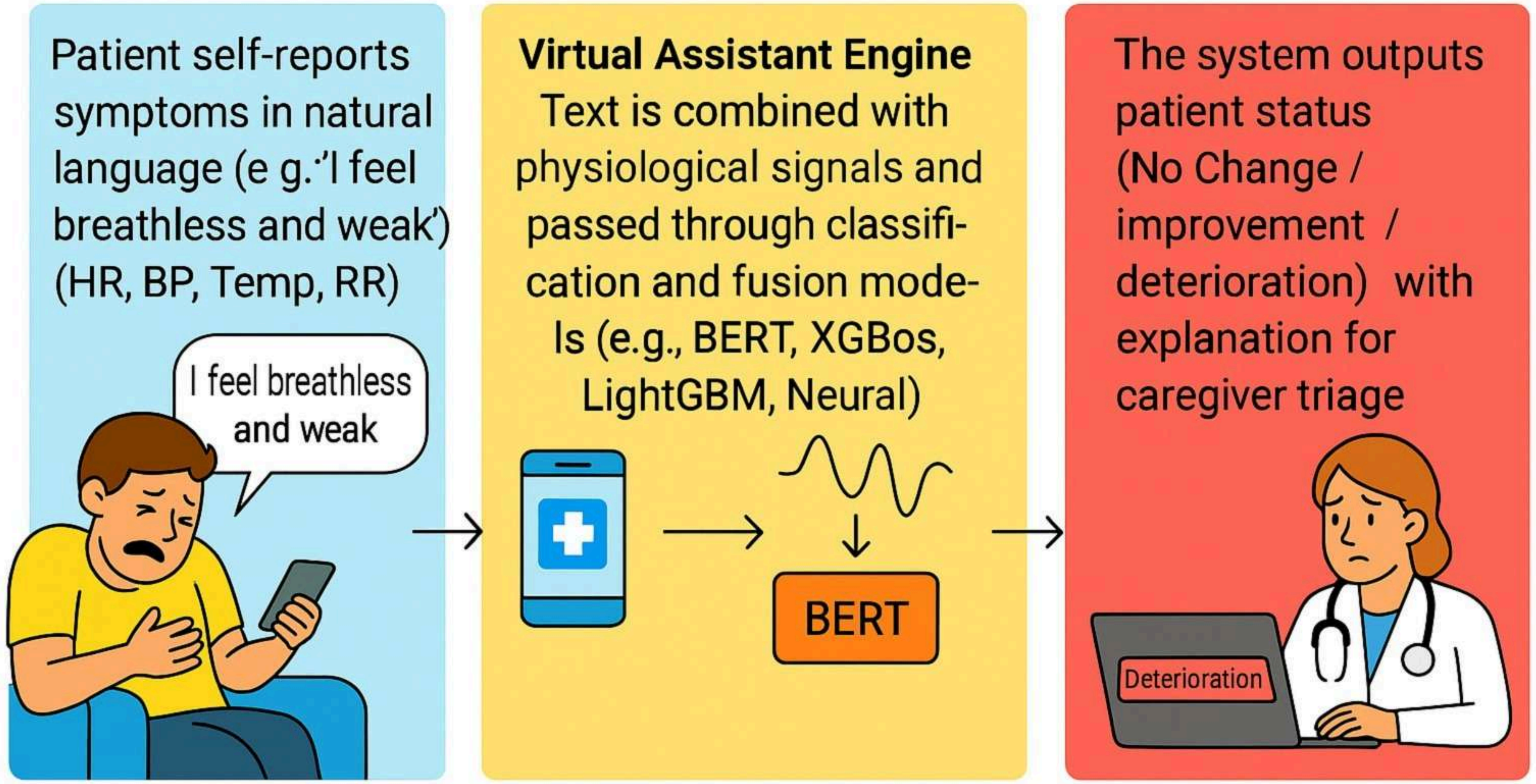
Accurate prediction of clinical change is achievable using accessible and interpretable models, as long as textual and clinical data are intelligently combined.



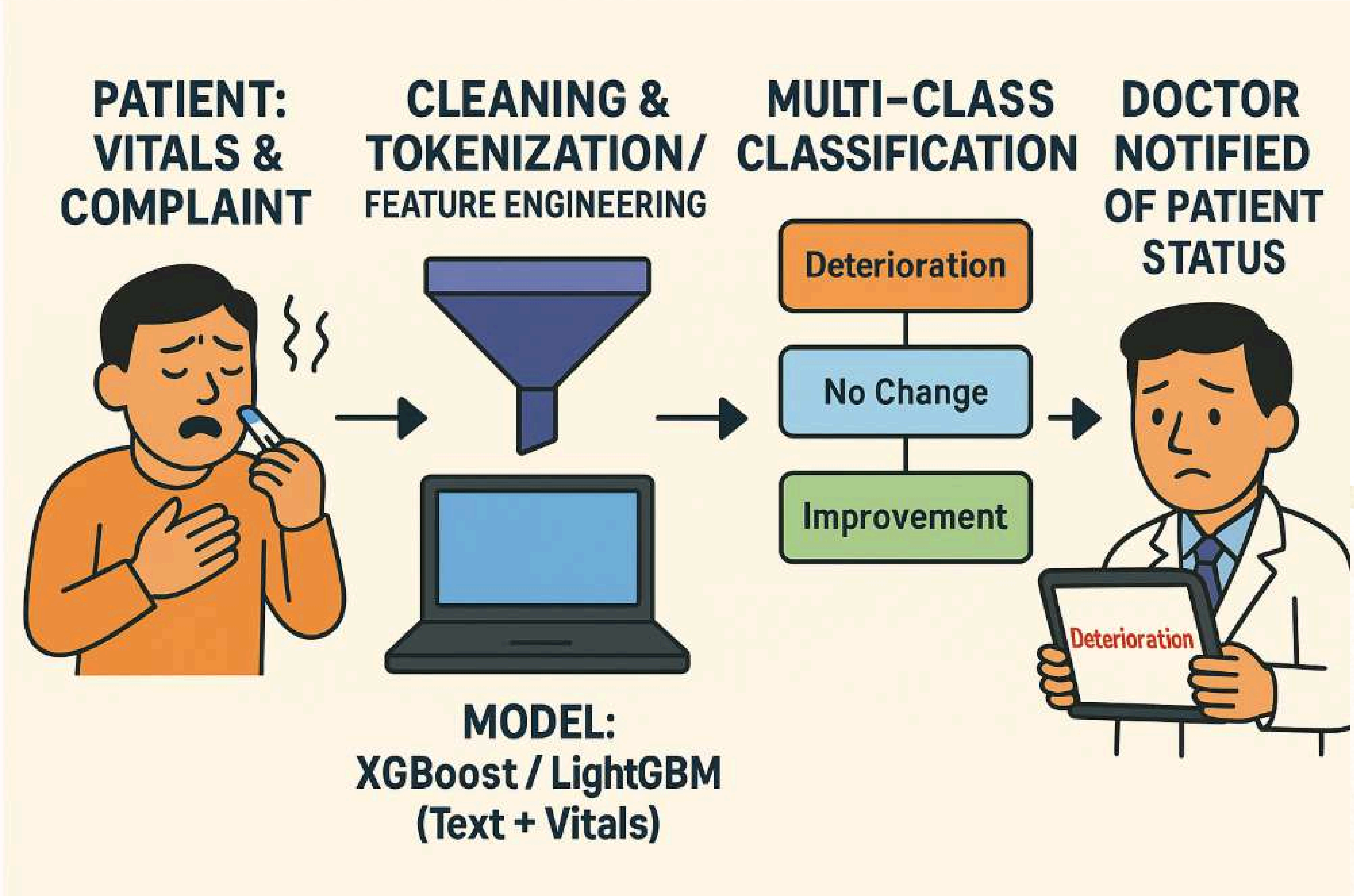


Graphical Abstract





Graphical Abstract



Thank You for listening