HORIZON 2020
The EU Framework Programme for Research and Innovation

CAPABLE
CAncer PAtients Better Life Experience

Grant Agreement No. 875052
Start Date: 01/01/2020 (48 Months)

Deliverable No. 5.7

Refined framework and models of all prototypes based on accumulated data

Due Date: [31/10/2023]
Submitted On: [20/11/2023]

Coordinator	University of Pavia (UNIPV)
Deliverable Lead Partner	IBM ISRAEL
Contributing Partners	IBM
Contact	Prof. Silvana Quaglini
Email	silvana.quaglini@unipv.it
Website	www.capable-project.eu

Deliverable Type		
R	Document, report	

This deliverable is a part of a project receiving funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 875052.

DEM	Demonstrator, pilot, prototype	**X**
DEC	Websites, patent filings, videos etc.	
OTHER		
	Dissemination Level	
PU	Public	**X**
CO	Confidential (Consortium members including the Commission Services)	
CI	Classified Information (Commission Decision 2015/444/EC)	

Table of Contents

List of Figures

List of Tables

1. Versions History

Version	Date	Author	Comments
1.0	15/10/2023	Simona Rabinovici-Cohen	Initial document
2.0	22/10/2023	Simona Rabinovici-Cohen, Itai Guez	Document before internal review
3.0	05/11/2023	Ronald Cornet	Internal Review
4.0	9/11/2023	Simona Rabinovici-Cohen, Itai Guez	Updates after Internal Review
5.0	16/11/2023	Silvana Quaglini	Final Review
6.0	20/11/2023	Simona Rabinovici-Cohen	Final Version

2. Executive Summary

The main goals of the statistical and predictions models in WP5 include clinical aspects, research aspects and technical aspects, as described below:

- **Clinical goal** – Enable more informed treatment selection and patient management by using automatic data-driven analysis models derived from CAPABLE clinical and sensors data. These models can support users of the Decision Support System (DSS) as well as of the Virtual Coach System that are part of the AI framework in CAPABLE architecture.
- **Research goal** – Advance the state-of-art methods and tools in predictive models and their interpretation for clinical practice. Specifically, we concentrate on multimodal models including imaging data that can benefit CAPABLE when imaging data will be accumulated in the future.
- **Technical goal** - Use and contribute to open-source frameworks and tools to enrich the biomedical research community and foster collaboration. This supports making technical assets developed within CAPABLE, sustainable beyond the project pilot.

The document is organized according to these three goals in the following way:

Section 3 provides a brief overview of the AI framework and its components, healthcare professionals (HCPs)' needs for statistical-based decision support, and the implementation of these needs by corresponding components in the AI framework.

Section 4 demonstrates the statistical analysis models performed on data collected from sensors (watches) provided to the CAPABLE pilot participants. It shows the correlation between the sensors time-series and the side effects that the patient encounters. The demonstration is presented as a video and includes all the steps in data analysis. Voice-over and a walk-through in this document explain all steps of the demonstration. This section supports the clinical goal of the statistical and predictions models in WP5.

Section 5 provides state-of-the-art multimodal models including clinical and imaging data to predict disease progression of kidney disease. This complements the prediction models done on the clinical data as described in previous deliverables and published paper (Barkan, 2022). It shows models that can be used in CAPABLE if additional multimodal imaging data is provided. This section supports the research goal of the statistical and predictions models in WP5.

Section 6 describes our contribution to open source, specifically to the BiomedSciAI GitHub organization. It shows the creation of sustainable assets and our increasing contribution to the biomedical community. This section supports the technical goal of the statistical and predictions models in WP5.

3. Statistics-based decision component as a part of the AI framework

The need for statistics-based decision components for HCPs and patients in the CAPABLE system was defined during the summarization of interviews and questionnaires in the Netherlands Cancer Institute (NKI) and in the Istituti Clinici Scientifici Maugeri (ICSM). There are two requirements involving statistics-based support (full details can be found in section 5 of D2.1 (Peleg, 2020)):

- HCPs are often involved in complex decision-making regarding treatment for patients. There is a need for supportive tools to facilitate this complex decision-making, for example, treatment choices, considering survival, and immune-related adverse events.
- Patients indicated varying levels of satisfaction with the information currently provided by their respective healthcare professionals. Patients are interested in a wide range of topics, including their diagnosis, treatments, and side effects.

The AI framework is part of the CAPABLE system architecture and was presented in D5.2 (Gilboa-Solomon, 2021). The aim of the AI framework is to provide support for both patients' and physicians' needs. The framework is defined as a set of concepts, libraries, tools, practices, and methodologies that cover formal knowledge representation, logic-based reasoning, and machine learning techniques.

The overview of the AI framework can be found in Figure 3.1**.** The components marked in red represent the WP5 components enhanced in previous deliverables while components marked in dark green line (in the bottom data-driven support) represent the specific components that were enhanced towards this deliverable. For this deliverable, we analyzed the pilot sensors data and created statistical population models that are described in section 4. We also analyzed UK Biobank data that includes large multimodal data with clinical, imaging and genomics information for patients with kidney disease. Using UK Biobank data, we created prediction models based on machine learning (ML) that predict kidney disease progression within five years. These state-of-the-art models can be used in CAPABLE when imaging data is available and may suggest to the physician and the patient how disease may progress and prepare for it. These ML-based prediction models are further described in section 5. To create these models, we utilized and contributed to BiomedSciAI, and specifically to FuseMedML open source, as further described in section 6.

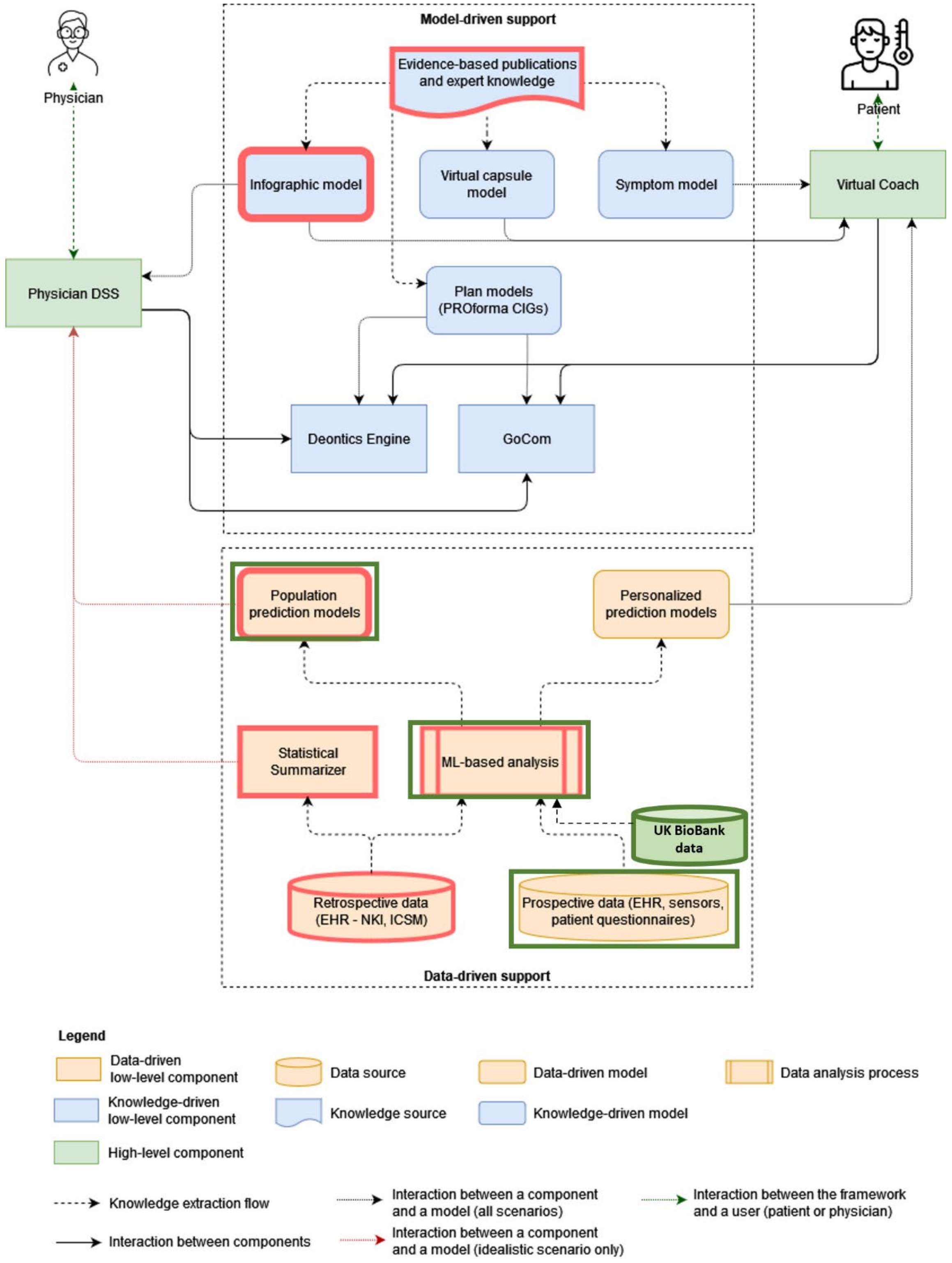
Model-driven support
Physician
Patient
Evidence-based publications and expert knowledge
Infographic model
Virtual capsule model
Symptom model
Virtual Coach
Physician DSS
Plan models (PROforma CIGs)
Deontics Engine
GoCom
Population prediction models
Personalized prediction models
Statistical Summarizer
ML-based analysis
UK BioBank data
Retrospective data (EHR - NKI, ICSM)
Prospective data (EHR, sensors, patient questionnaires)
Data-driven support
Legend
Data-driven low-level component
Data source
Data-driven model
Data analysis process
Knowledge-driven low-level component
Knowledge source
Knowledge-driven model
High-level component
Knowledge extraction flow
Interaction between components
Interaction between a component and a model (all scenarios)
Interaction between a component and a model (idealistic scenario only)
Interaction between the framework and a user (patient or physician)

Figure 3.1: Overview of the AI framework and the new contributions. Components marked in dark green lines (in the bottom data-driven support) represent the specific components that we enhanced towards this deliverable.

4. Demonstration of data-driven statistical analysis of sensors data

4.1. Overview

To make the CAPABLE dashboard and app more informed and data-driven, it should be useful to analyze sensors data coming from watches provided to the patients of the CAPABLE pilot, specifically data such as SPO2 levels, blood pressure measurements, activity measurements and more. This could provide additional clinical decision aids and educational aids for discussing the diagnosis, treatments, prognosis, and side-effects if the smartwatch was a medical device. Unfortunately, up to date, the ASUS device provided to patients is not a medical device and, as such, it cannot be currently used to suggest actions to patients or doctors.

As a matter of fact, it seems that the ASUS VivoWatch 5 purpose is to monitor a healthy person's heart metrics and training activity while doing sports rather than measuring illnesses and diagnosis of severe chronic diseases and their symptoms. As mentioned, the watch is not a CE-certified medical device and thus ASUS manufacturer notes the following in its documentation:

(see https://www.asus.com/mobile-handhelds/wearable-healthcare/asus-vivowatch/asus-vivowatch-5-hc-b05/)

• ASUS VivoWatch 5 is not a medical device and is not intended for diagnosing medical conditions. Measurement results are for reference only.

• Do not adjust medication based on the measurement results from ASUS VivoWatch 5. Take medication as prescribed by your physician. Only a physician is qualified to diagnose and treat high blood pressure.

Our recommendation is to use CE-certified devices for medical sensor data collection,

such as Tyto (https://www.tytocare.com/products/tytohome).

Thus, as a very important premise, we inform the reader that the analyses performed and described in the following have not been integrated into the CAPABLE system deployed at the hospital. The aim of the analysis has thus been purely research related.

We analyzed the sensors time-series for ICSM and NKI pilots including measurement of anomalies and their correlation with side effects. We created a demonstration of all the steps of our analysis that can be found at:

https://capable-project.eu/wp-content/uploads/2023/11/D5.7_Demo.mp4

4.2. Data sources

In this deliverable development period, we assigned ASUS smartwatch VivoWatch5 to collect essential time-series data during normal activity of the patients in the CAPABLE pilot. We collected sensors data from two pilots as described below. Note that the CAPABLE pilot is ongoing, and our analysis refers to data collected till October 15th, 2023.

- ICSM pilot
    - 53 watches were provided to ICSM patients.

  - 2 patients dropped out.
  - 4 watches were assigned for test.
  - 3 watches did not gather any data until now.
  - Total 44 watches collected data from real patients.
- NKI pilot
  - 27 watches were provided to NKI patients.
  - 1 patient dropped out due to a rash on the wrist.
  - 3 watches were not given to patients.
  - 3 watches did not gather any data until now.
  - Total 21 watches collected data from real patients.

Each watch has a battery life for about 10-14 days of usage according to the tech specs mentioned by the ASUS manufacturer (https://www.asus.com/mobile-handhelds/wearable-healthcare/asus-vivowatch/asus-vivowatch-5-hc-b05/techspec/). After that period, or probably earlier according to the usage intensity, the watch needs to be charged and will not collect patient data during charging.

Each watch included the following three sensors to collect data:

- G sensor - accelerometer and ambient temperature sensor
- PPG (photoplethysmography) and ECG (electrocardiography) sensors
- Built-in GPS sensor

Each sensor can be turned off or on by the user. The watch reports sensor data once per minute when activated.

The collected data from the sensors includes basic measurements such as:

- Ambient temperature in Celsius
- SPO2 level in blood
- Estimate of the Blood pressure (BP) measures, systolic blood pressure (sys), diastolic blood pressure (dia), based on PPG and EEG
- Heart Rate (HR)
- Heart rate variability (HRV) measures - SDNN (standard deviation of intervals), RMSSD (root mean square of successive interval differences), LFP (low frequency power), HFP (high frequency power), and Stress (total score based on several HRV measures). Power is the signal energy found within a frequency band.
- Activity – steps, calories, toss & turn
- Sleep state – light, awake, deep, rapid eye movement (REM) sleep

The smartwatch extracts the heart related features from ECG and PPG raw signals. We noticed that measures based on ECG or PPG alone are less accurate than considering both signals together. In fact, this is also documented in the ASUS manual. Thus, we considered only measures calculated using both ECG and PPG and extracted features from them for our analyses.

4.3. Demonstration of statistical analysis

In this subsection, we illustrate the statistical analysis of collected sensors data to both analyze the quality of the sensors themselves and to analyze the benefits of the sensors in identifying clinically meaningful events e.g., worsening of medical conditions, or events that need urgent treatment such as life-threatening events.

After first inspection of the sensor values, it seems that some sensors produced unrealistic values by any known measure of human beings (e.g., heart rate of 1 and 248, SPO2 levels higher than 100%). These values were defined as anomalies and were filtered out e.g., by setting the range for acceptable values of heart rate between 40-200.

When working with filtered values, anomalies were defined on the following numerical features. Heart Rate Variability (HRV) anomalies were not defined as there are no acceptable criteria for values which represent a medical condition and are in academical consensus. Normal ranges and non-outlier were defined based on evidence in literature and feature definition.

Table 4.3.1: Watch features and their normal and non-outlier range

Feature	Normal range	Non outlier range
BP_hr	**60-100**	**40-200**
BP_sys	**<140**	**>=0**
BP_dia	**<90**	**>=0**
SPO2_value	**95-100**	**0-100**

Next, patients/watches were ranked by the percentage of anomalous records they have out of the entire number of records per feature per watch. Watches with top anomalies in ICSM pilot are attached in the table below. The table shows the top data anomalies found, in which a feature is more than 30% of the time outside its normal range. Each major anomaly can be attributed to either:

- Technical issue with the watch
- Usage not according to watch manufacturer – either its positioning or proper wearing
- Bug in the ASUS watch firmware/hardware
- Environmental change (going outside, doing activity, charging the watch)
- Real medical anomaly measurement of the patient wearing the watch

Careful manual investigations of the watches and patient history is needed to understand what the cause of these anomalies is.

Table 4.3.2: Watches with top anomalies percentage in ICSM pilot

watch	feature	min	max	anomalies	total	percent
VivoWatch5-CAPABLE-0029	BP_hr	52	119	8	12	67%
VivoWatch5-CAPABLE-0033	SPO2_value	1	121	654	1148	57%
VivoWatch5-CAPABLE-0028	SPO2_value	1	114	1345	2668	50%
VivoWatch5-CAPABLE-0029	SPO2_value	70	99	1049	2311	45%
VivoWatch5-CAPABLE-0022	BP_dia	0	249	4	9	44%
VivoWatch5-CAPABLE-0022	BP_sys	0	159	4	9	44%
VivoWatch5-CAPABLE-0053	SPO2_value	2	111	200	452	44%
VivoWatch5-CAPABLE-0034	SPO2_value	9	123	3148	7709	41%
VivoWatch5-CAPABLE-0022	SPO2_value	1	127	5080	12580	40%
VivoWatch5-CAPABLE-0030	BP_dia	1	104	74	185	40%
VivoWatch5-CAPABLE-0025	SPO2_value	3	115	327	838	39%
VivoWatch5-CAPABLE-0001	BP_hr	47	89	3	8	38%
VivoWatch5-CAPABLE-0003	SPO2_value	1	110	871	2379	37%
VivoWatch5-CAPABLE-0001	SPO2_value	1	126	1633	4779	34%
VivoWatch5-CAPABLE-0020	SPO2_value	1	123	409	1203	34%
VivoWatch5-CAPABLE-0047	SPO2_value	1	125	1554	4712	33%
VivoWatch5-CAPABLE-0027	SPO2_value	1	127	2456	7784	32%
VivoWatch5-CAPABLE-0001	BP_sys	0	157	3	10	30%

Next, we tried to find correlations between the side effects that the patients reported in the clinical data and the sensor data anomalies found in the watches (after filtering out measurements errors). The table below shows the side effects that the patients reported in the ICSM pilot while wearing their watches. Note that no cardiovascular life-risking events were reported by the patients.

Table 4.3.3: Reported side effects during watch usage in ICSM pilot

Event	Num patients	Patient's watches
Backache	6	41, 31, 29, 26, 30, 49
Cough	5	29, 6, 17, 5, 20
Diarrhea	13	3, 1, 31, 11, 25, 21, 5, 36, 20, 30, 34, 35, 54
Fatigue	7	22, 28, 5, 27, 36, 35, 45
Fever	5	3, 11, 33, 20, 30
Headache	9	1, 9, 14, 26, 30, 36, 46, 47, 49
Insomnia	2	36, 54
Nausea	12	1, 3, 5, 20, 21, 25, 27, 30, 36, 31, 46, 54

Regarding fatigue side effects, we found concurrency between fatigue reports and lower than normal SPO2 percentage more than 30% of the time on watches 22, 27, and 28. We saw that SPO2 sensor measurements are at a low level (most of the time below 95%) when fatigue side effects are reported. The figure below includes the SPO2 graph of watch 28 which has lower levels of SPO2 sometimes when fatigue is present. This patient reported fatigue side effects during the entire period (16/5/23-12/7/23). Additional plots of the entire time series of these 3 watches 22, 27, and 28 appear in Annex 9.1.

When counting all the watches, the correlation between fatigue and low SPO2 was not proved to be statistically significant. The intersection over union (IOU) for all the watches is 0.21. The Spearman correlation gave 0.08 (p-value=0.66). The Pearson correlation also gave the same results of 0.08 (p-value=0.66).

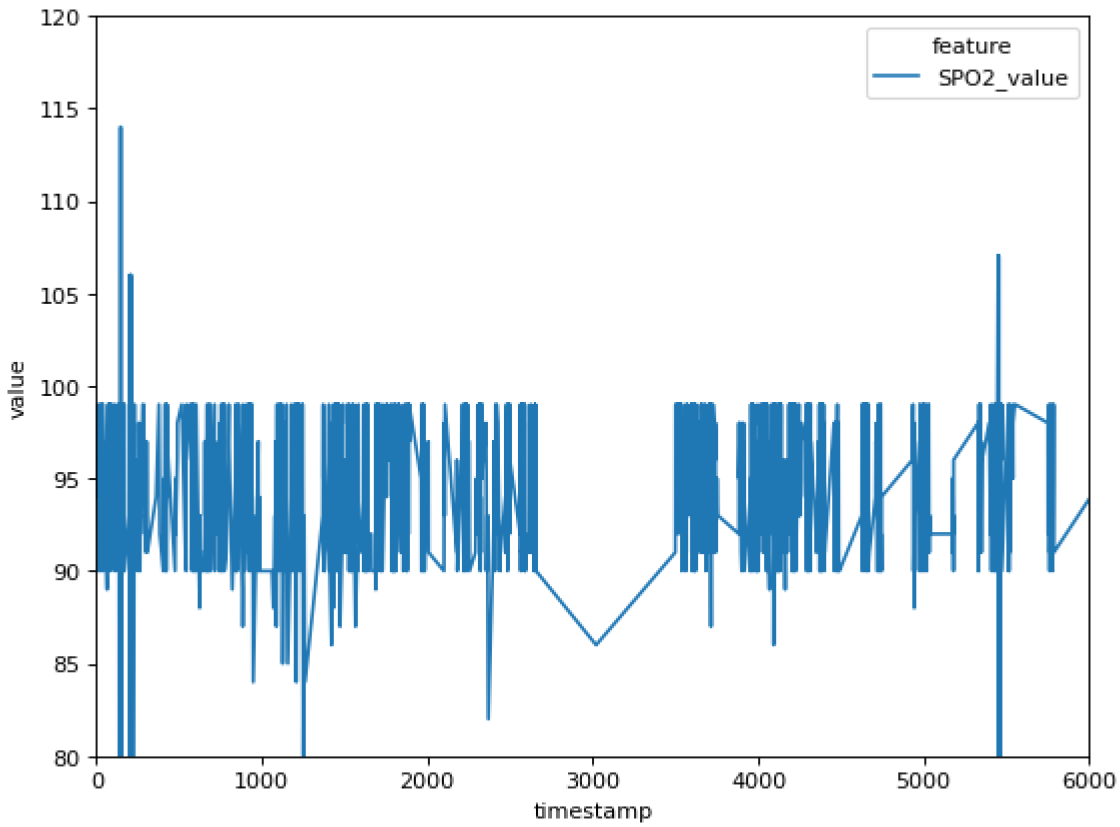

Figure 4.3: Watch 28 with SPO2 anomaly concurrent with fatigue side effect

The following table presents the top anomalies found in the NKI pilot study in which a feature is more than 30% of the time outside the normal range. For this pilot, we didn't have the reported side effects, as we didn't get the permissions to see them.

Table 4.3.4: Watches with top anomalies percentage in NKI pilot

Watch	feature	min	max	anomalies	total	percent
VivoWatch5-CAPABLE-NKI-011	BP_dia	193	193	1	1	100%
VivoWatch5-CAPABLE-NKI-005	SPO2_value	1	120	191	277	69%
VivoWatch5-CAPABLE-NKI-011	SPO2_value	1	126	226	439	51%
VivoWatch5-CAPABLE-NKI-013	SPO2_value	92	97	2	4	50%
VivoWatch5-CAPABLE-NKI-009	BP_sys	127	142	1	2	50%

VivoWatch5-CAPABLE-NKI-008	SPO2_value	79	99	62	148	42%
VivoWatch5-CAPABLE-NKI-019	SPO2_value	90	99	42	136	31%

We also analyzed the activity including total number of steps and calories during the period wearing the watch. The results of this analysis for both ICSM pilot and NKI pilot appear in Annex 9.2. We then calculated correlation between activity and heart rate as well as correlation between activity and SPO2.

4.4. Results and Conclusions

In this section we will summarize the functionality and benefits of ASUS VivoWatch 5 in general, and review in specifics each of the watch sensors in terms of which sensor data is useful or not useful for our analysis. We will consider the usefulness of each watch sensor separately.

4.4.1. Temperature

The ambient temperature sensor in the ASUS watch is not accurate to measure a patient core temperature and fever, as in many watches the temperature was irregular for more than half the time while no fever was reported. This could be explained by the fact that the temperature is an ambient temperature sensor – which means it measures the temperature outside the body rather than the patient’s core temperature. Therefore, the sensor is affected by the temperature outside, the watch hardware temperature, and the patient’s skin temperature.

The figure below shows the temperature measurements over time. It seems that the temperature sensor measurements oscillate all the time with a standard deviation which is higher than needed to reliably measure a temperature. Therefore, the temperature goes beyond the normal range even though the patient is not sick.

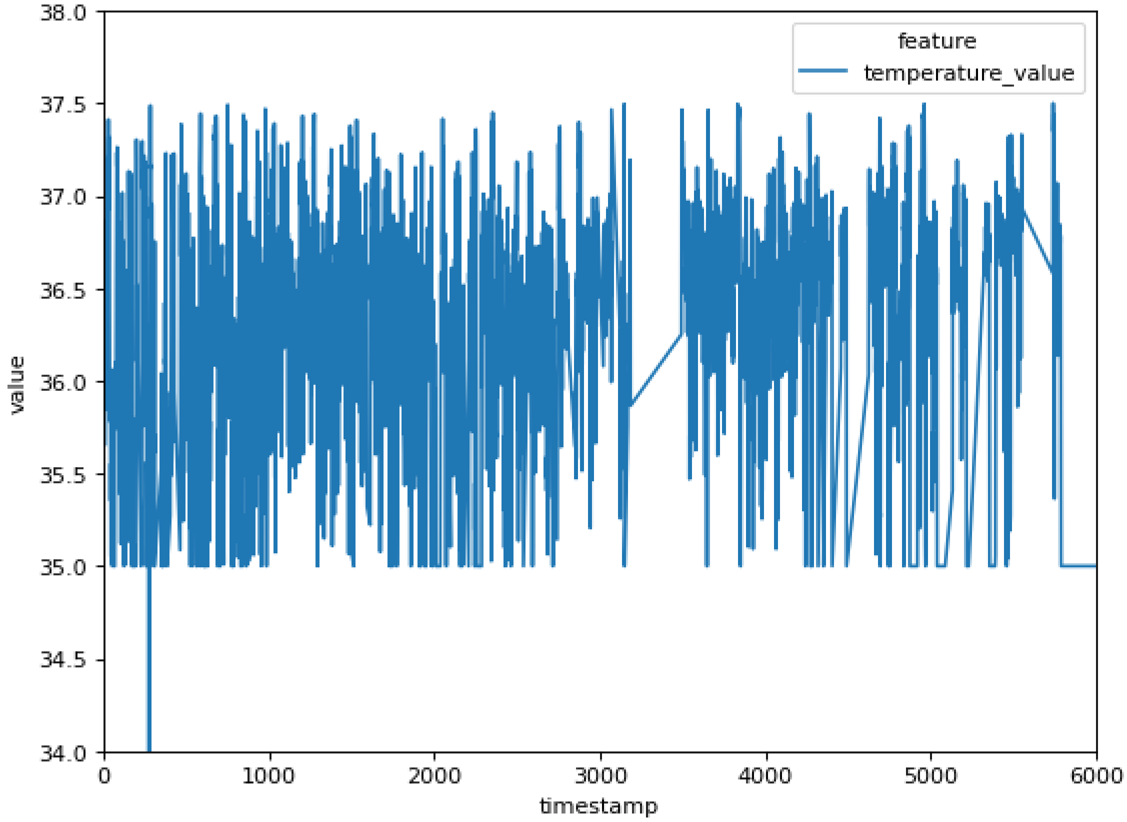

Figure 4.4: Temperature sensor over time

In other cases, the temperature data is sparse and has outliers which cannot be explained by the clinical data, but rather by change in the environment (charging, going outside).

In conclusion, the ASUS watch ambient temperature is useless for the purpose of measuring a patient's core temperature.

4.4.2. SPO2 oxygen saturation

SPO2 measurements are not always accurate. However, we have seen that for 3 watches, SPO2 levels that go 30% or more of the time below the normal range were correlated with reports of fatigue. For other watches, either patients wearing them did not report fatigue or they reported fatigue but had normal SPO2 levels for more than 70% of the time (4 watches).

In conclusion, the ASUS SPO2 level feature might be useful to identify fatigue events. However, more data needs to be gathered and analyzed regarding SPO2 level drop and side effects, to estimate its preciseness in terms of specificity/sensitivity for each clinical target.

4.4.3. Activity correlation with heart rate and SPO2

Getting a signal from ECG or PPG solely is not reliable enough. When using both ECG and PPG measurements that are taken intentionally by the patient, the data points become scarce but accurate. Some watches don’t have such measurements or don’t have more than 1-3 measurements a day.

Correlation was calculated between activity (steps per day) and heart rate and SPO2. A negative correlation was found between activity and heart rate, and a positive correlation was found between activity and SPO2. These findings are in line with what is known from clinical practice. For ICSM: corr(activity, hr) = -0.2, p-value=0.33. For ICSM: corr(activity, SPO2) = 0.25, p-value=0.27. For NKI: corr(activity, hr) = -0.04, p-value=0.94. For NKI: corr(activity, SPO2) = 0.38, p-value=0.52. The correlation is not considered strong and therefore the relation is not close to linear. However, its sign has more information than the absolute value, as we don’t necessarily seek a linear relation but rather the trend of any kind of function between these two features which the linear correlation test approximates.

In conclusion, the activity measurement sensors reliably measure the patient activity while wearing the watch. It is useful for tracking the patient's fitness and daily behavior, which might be related to its current general health status, fitness, and mood.

4.4.4. Heart rate, blood pressure and heart rate variability

Heart rate and blood pressure sensor measurement are theoretically more exact when the patient initiates the measurement and activates both ECG and PPG sensors together for signal extraction. However, the number of such samples in our data is sparse and more cardiovascular clinical events or alternative devices for comparison are needed to conclude the exact utility and precision of those sensors.

HRV is an emerging and relatively new field of research in the study of ECG and PPG heart signals. Its clinical significance and usefulness as a biomarker in measuring no life risking events, is not in academic consensus yet and therefore was not used in this report.

5. Multimodal prediction models for disease progression

5.1. Overview

Prediction models produce findings from the analysis that can be further applied to new patients to provide insights related to clinical treatment, such as disease progression, response to treatment, survival, and toxicity.

Multimodal prediction models are considered state-of-the-art models that provide better performance than unimodal models (Rabinovici-Cohen, 2022). In previous deliverables, we explored unimodal prediction models learned from CAPABLE retrospective clinical data and published the results in a Frontiers paper (Barkan, 2023). Towards this deliverable, we wanted to explore the contribution of multimodal models, and to investigate its potential benefits to CAPABLE when imaging data will be collected.

We trained our models on UK Biobank which is a source of large multimodal data, and built models to predict the disease progression of chronic kidney disease (CKD). There were not many patients in UK Biobank with kidney cancer and multimodal data, so we chose CKD because it had many patients with multimodal data, and it relates to the ICSM pilot. As a matter of fact, there is a strong and bidirectional relationship between CKD and renal cancer, as illustrated in (Saly 2021), where we read “Chronic kidney disease can lead to the development of renal cell carcinoma via oxidative stress from a uremic milieu or an underlying cystic disease. Surgical management of renal cell carcinoma can lead to chronic kidney disease via reduced nephron mass or acute kidney injury events. Medical management of renal cell carcinoma can lead to acute kidney injury, which can lead to chronic kidney disease.”

A paper that summarizes our work on imaging and clinical data as well as additional work on genomics data is in-progress.

5.2. Data sources

We trained our models on data from UK Biobank (https://www.ukbiobank.ac.uk/) using Neck-to-Knee imaging scans as well as the clinical data for the same patients. The UK Biobank has data from about half a million patients; from which about 50,000 patients have magnetic resonance imaging (MRI).

The MRI data includes scans taken with a Siemens Aera 1.5T device that acquired overlapping images in six stations covering the body from neck to knee. Each station has low resolution imaging of size 224x174x44 pixels. The kidneys are typically located in the second and third imaging stations. There are four types of sequences for each station depending on how the image scanning is configured, and these sequences are: water only (water), fat only (fat), in-phase (in) and opposed-phase (opp).

The clinical data includes demographics data: age, gender as well as Clinical Classifications Software (CCS) diagnosis codes from ICD-9/ ICD-10 that represent the multiple diagnosis that the patient has and the patient’s clinical condition.

5.3. Predictive models for chronic kidney disease

In this subsection, we describe the study design and the various methods we used.

5.3.1. Study Design and Patients Data

Chronic kidney disease (CKD) is a condition where the kidneys are damaged and progressively lose their ability to filter blood. Generally, CKD patients progress over multiple CKD stages, often slowly and heterogeneously, from no disease (CKD 0) to mild kidney damage (CKD stages 1 or 2) to severe kidney damage (CKD stages 3 or 4 or 5) to kidney failure and need for dialysis (CKD dial).

In our work, we concentrated on two prediction tasks:

- Prediction Task 1: Predict whether the patient with no CKD (CKD 0) will progress to severe CKD (CKD 3, 4, 5, dial) within five years.
- Prediction Task 2: Predict whether the patient with no/light CKD (CKD 0, 1, 2) will progress to severe CKD (CKD 3, 4, 5, dial) within five years.

We got similar results for Task 1 and Task 2, so we'll mainly describe task 2. Annex 9.3 describes some of the results for task 1.

The figure below depicts the study design. We analyzed multimodal data including clinical and imaging data. Another team that we collaborated with also analyzed genomics data. The index date, namely the date when the prediction is done, is the date in which the imaging was taken. We also consider all the clinical data available at that index date: age, gender, and diagnosis codes that the patient had by that time.

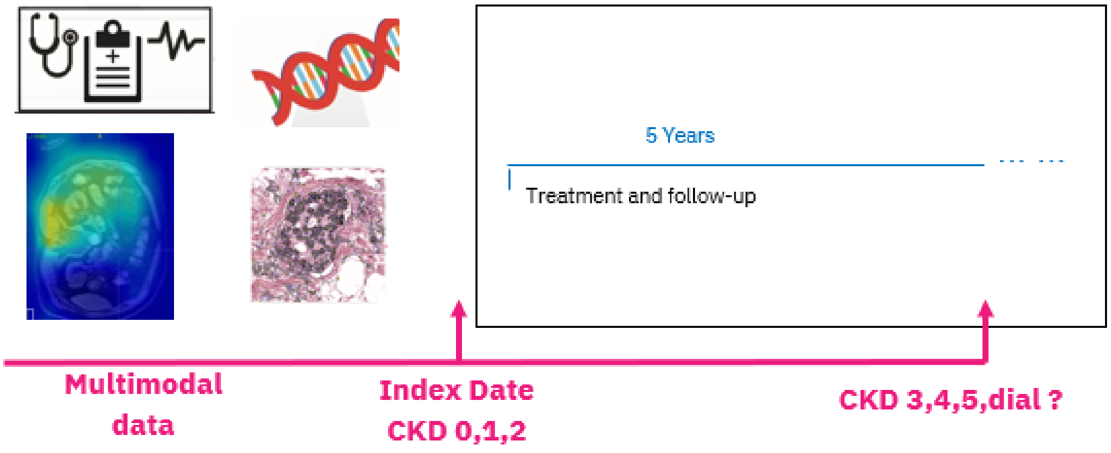

Figure 5.3.1: Study design

5.3.2. Methods for the prediction models

We split the clinical and MRI data into 5 folds and we performed 5-fold cross-validation. In that process, we iteratively selected a different validation fold and trained on the remaining

4 folds, resulting in 5 models validated on different folds. We then average the five models selected from the five-fold cross-validation.

The overall method is depicted in the figure below. We create models for each modality alone and then ensembled the multiple models. We have five unimodal models: (1) a clinical model based on age and gender (2) a clinical model based on the CCS codes (3) an imaging model based on radiomics (4) an imaging model based on convolutional neural networks (CNN) (5) an imaging model based on visual transformer (ViT). For each model we extract features and then apply ML classifiers (XGBoost, Random Forest, Logistic Regression). We select the classifier that gives the best performance for selected features.

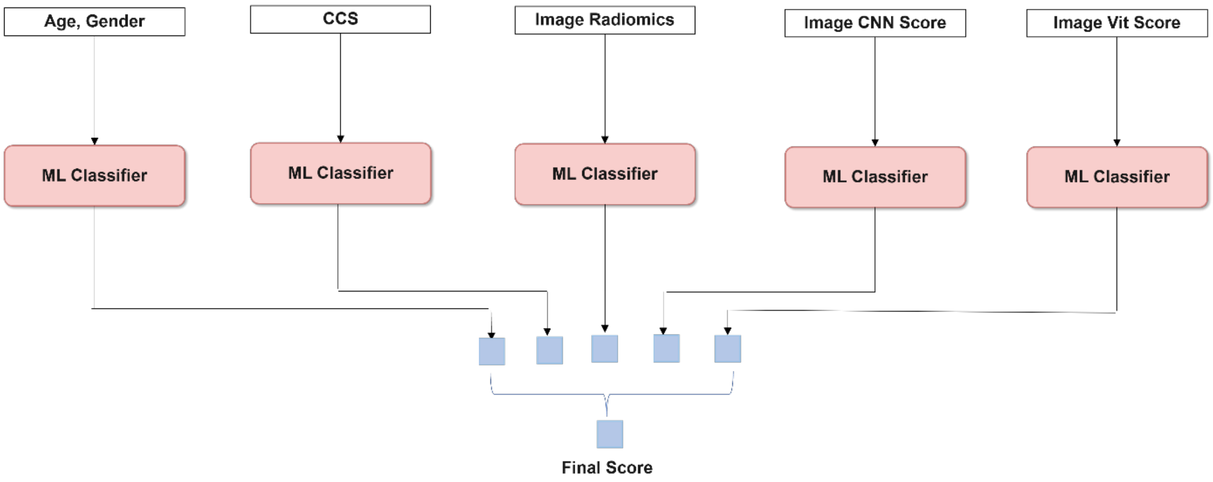

Figure 5.3.2: Overall method

As mentioned, there are three different imaging methods to analyze our imaging data: Radiomics, CNN and ViT. Each method provides a different set of features and thus enriches our overall model. Figure 5.3.3 below summarizes the three imaging methods.

Radiomics is a quantitative approach to medical imaging. Its goal is to find associations between qualitative and quantitative information extracted from clinical images and clinical data by using analysis methods from the field of computer vision, information theory and statistics. For any radiomic approach, it is critical to define the volume of interest (VOI) in a three-dimensional (3D) volume from which the radiomics features will be calculated. Kidney 3D segmentation to be used as VOI to extract radiomics were generated by a pretrained segmentation model (2.5D U-net) from previous research (Langner, 2020). Using the VOI, we then extracted all the supported features in the pyradiomics python package.

For the CNN model, we used Video-resnet CNN (Tran, 2018) and adapted it to the kidney MRI that we had. Our image preprocessing included isotropic pixel-spacing, intensity range normalization, and discretization.

For the Vit model, we adapted the vision transformer (Dosovitskiy, 2020) to our kidney MRI data. We divided the data into consecutive patches that were then fed into the transformer followed by a multi-layer perceptome (MLP) to predict the score of the disease progression.

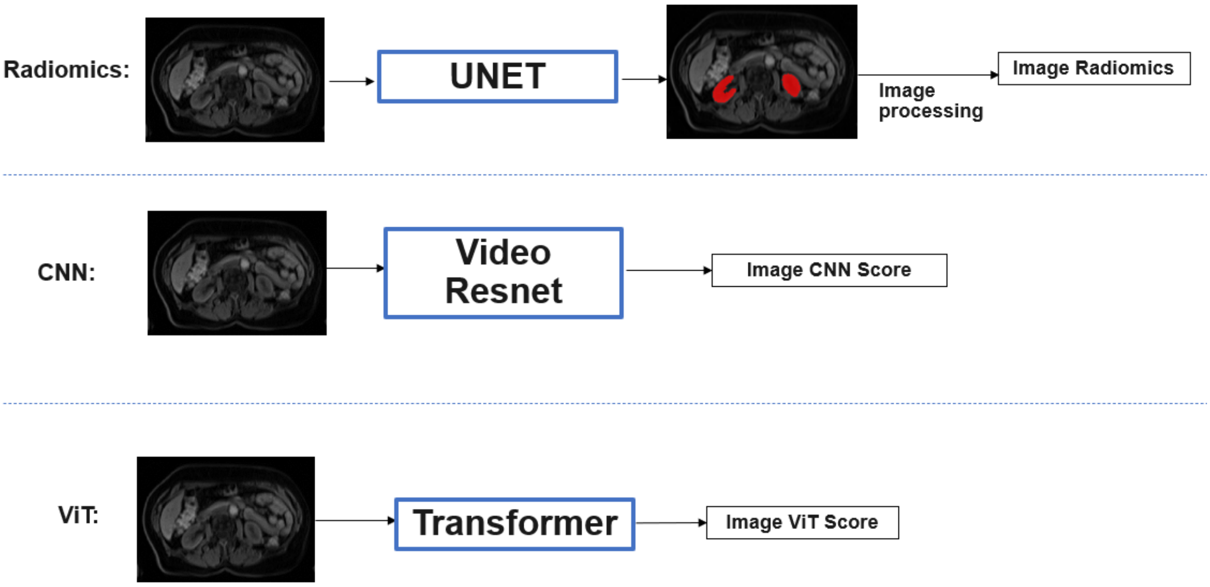

Figure 5.3.3: MRI imaging method

The ensemble model receives five scores per patient: two scores based on clinical data and three scores based on the MRI data. To improve generalization, we used different training seed initialization for the different models. We then examined several strategies for combining and 'ensembling' the models. However, we found that the most effective strategy used the mean value of all available scores per patient.

5.3.3. Models' evaluation and explanation

The radiomics model gave the best unimodal results. The figure below shows the ROC curve for the five-fold cross-validation of the radiomics model. Logistic regression was found to be the best classifier for this kind of features. When averaging over the five cross-validation models, we get 0.745 area under the curve (AUC).

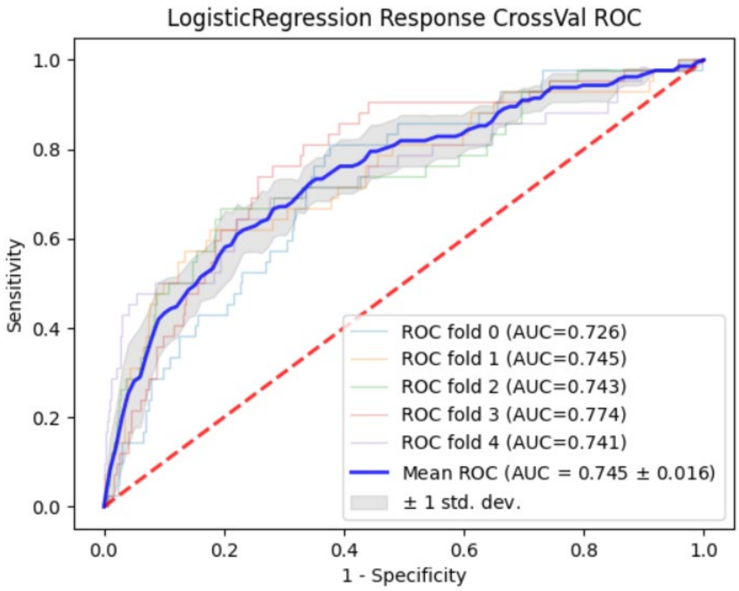

Figure 5.3.4: Radiomics model five-fold cross-validation ROC curve

The results of the other unimodal models are in Annex 9.4.

We also performed SHAP (Lundberg, 2017) analysis to explain the features that contributed to the radiomics model. SHAP considers all possible combinations of features with and without a specific feature to evaluate its contribution to the prediction. It reveals each feature importance and demonstrates how each feature of each patient affects the predictive model results. The figure below depicts the top radiomics features in descending order that had the most influence on the severe CKD prediction. A positive SHAP value for a feature means it leads the model to predict 'severe-ckd', while a negative value leads the model to predict 'no-severe-ckd'. The point's color represents the values that each feature can take, including red for high values, blue for low values, and purple for values that are close to the average value.

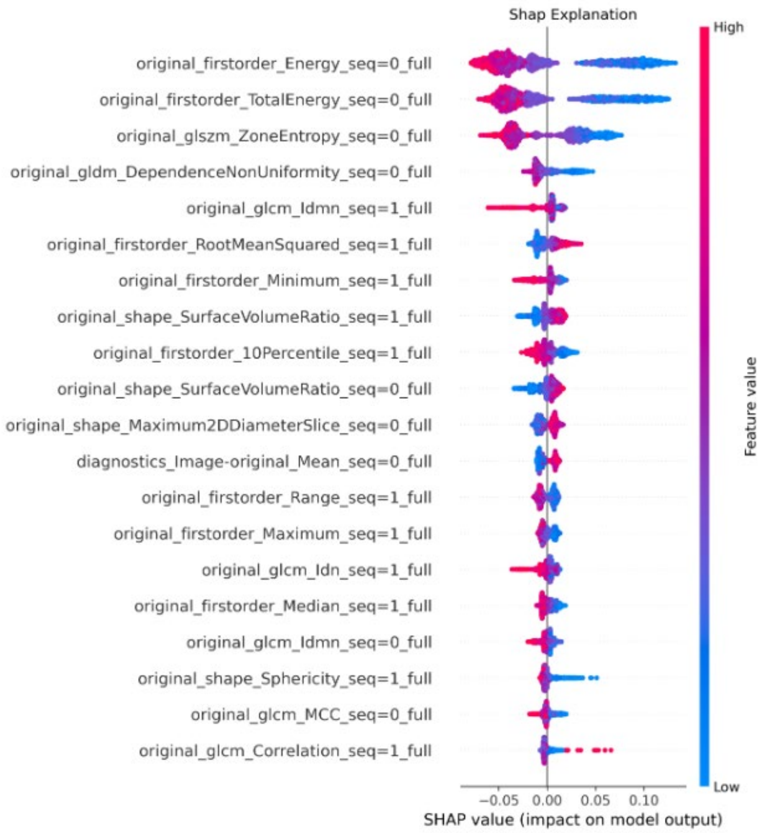

Figure 5.3.5: SHAP explanation for the Radiomics model

As mentioned, each MRI has four sequences: water (seq=0), fat (seq=1), in (seq=2) and opp (seq=3). We note that the most important features are from the water and fat sequences. This means that for our purposes, there is no need to do the extra two sequences in (seq=2) and opp (seq=3). Performing less scans can benefit the patient and reduce costs.

The next figure shows the result of the ensemble model. We see that the multimodal ensemble model that combines all the unimodal models is better than each unimodal model alone and achieves 0.804 AUC.

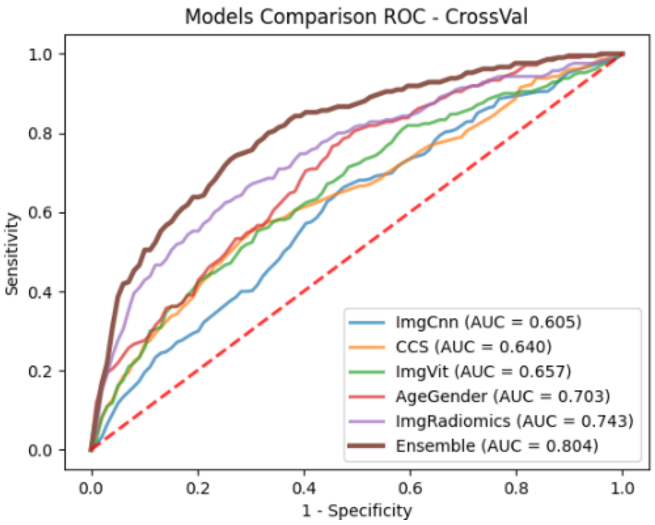

Figure 5.3.6: Ensemble model ROC curve

We also used the Delong test (Delong, 1988) to calculate the p-value when comparing the prediction of the individual models with those of the ensemble model. We received that the ensemble model is statistically better (p-value < 0.05) than each modality model alone.

5.4. Conclusions

Chronic kidney disease is a dynamic disease and making an accurate prediction whether the disease will progress within five years is challenging. Accurately predicting the future disease progression based on data available prior to treatment initiation could impact the treatment planning and selection. We have rich information collected from the UK Biobank including clinical information and MRI medical imaging. We introduce a multimodal prediction model that is based on clinical data and neck-to-knee MRI images taken prior to disease progression. We compared performance of models on different data elements and evaluated them by AUC. The results on cross-validation show that the multimodal ensemble model that leverages both the MRI and the clinical models offers improved results over the unimodal models. We then used interpretability methods to explain the model and identify important features for predicting disease progression.

The results we obtained on the UK Biobank data can be used in the CAPABLE system when additional imaging data is aggregated. Recent state-of-the-art research shows that large foundation models can serve as pretrained models for new tasks achieving impressive results with only small data for the new task fine-tuning (Moor, 2023). This suggests that the models we created using the UK Biobank data can serve as pretrained models for CAPABLE and further fine-tuned for the CAPABLE tasks even with a small multimodal dataset.

6. Open-source contribution to BiomedSciAI open GitHub organization

6.1. Overview

Artificial Intelligence (AI) is leading the way in scientific advancements in biomedicine, holding the promise to enhance our lives and reduce healthcare costs. To accelerate scientific discovery in this intricate field, it is crucial to have frameworks and tools that foster collaboration among researchers and developers, enabling them to collaborate, reuse components, and reproduce results.

The BiomedSciAI GitHub organization (https://github.com/BiomedSciAI) includes open-source repositories to foster scientific discovery in biomedicine. It includes flexible frameworks and tools designed for easy collaboration, encouraging code reuse. It received significant recognition, reflected in over 950 GitHub stars, 183 forks, and close to 130,000 downloads. The table below shows the various repositories in BiomedSciAI with various topics and modalities including clinical, imaging, pathology, genomics and more.

Table 6.1: BiomedSciAI GitHub organization and its open-source repositories

GIT Repository	Description
Causallib	causal inference analysis
FuseMedML	deep learning framework
Histocartography	pathology data analysis
DPM 360	clinical data analysis
fuse-drug	for drug discovery
r-BRICS	breaking chemical substructures
MMMT	Multi-modal models toolkit
Geno4SD	genomics data analysis

As part of CAPABLE we reused and contributed to BiomedSciAI and especially to the FuseMedML repository that is further described in the next subsection. For example, we contributed the EHR transformer example - https://github.com/BiomedSciAI/fuse-med-ml/tree/master/fuse_examples/multimodality/ehr_transformer. This example shows using a foundation model with FuseMedML to analyze EHR data. The foundation model was built with a transformer (BERT) that can model long-range dependencies and capture complex patterns in sequential data such as the biomedical data.

6.2. FuseMedML open source

FuseMedML is an open-source project (https://github.com/BiomedSciAI/fuse-med-ml) released under the Apache-2.0 license and has dozens of contributors over the span of two years. It is part of the PyTorch Ecosystem (https://pytorch.org/ecosystem/), which is a large community with over 80 open-source projects including projects with thousands of stars and millions of downloads.

FuseMedML employs a structured architecture with decoupled components that can be reused independently and thus enable easy adoption and low entry barrier. The core code is based on many popular open-source projects such as scikit-learn, PyTorch and PyTorch lightning. Furthermore, fuse comes with a rich collection of modular domain specific implemented components. This modularity also enables easy extension of the framework with additional functionally and to additional domains.

Fuse is structured in three layers as depicted in the figure below. The bottom layer includes standalone basic components that can be reused independently by the other components. This includes fuse.data, a flexible data processing pipeline with functionalities such as augmentations and caching. The other standalone component is fuse.eval, a library for evaluating AI models including various metrics and methods for model comparison. The middle layer uses the bottom layer and consists of fuse.dl with implemented reusable deep learning components such as data loaders, backbone models and loss functions.

The core technology of FuseMedML and its component packages is general, while domain specific functionality is contained in the top layer within extensions. These include fuseimg that extends the data package for processing of medical imaging from various modalities, fusedrug for therapeutic molecules generation, drug discovery and repurposing, and fuse_examples, a rich set of end-to-end examples based on open data.

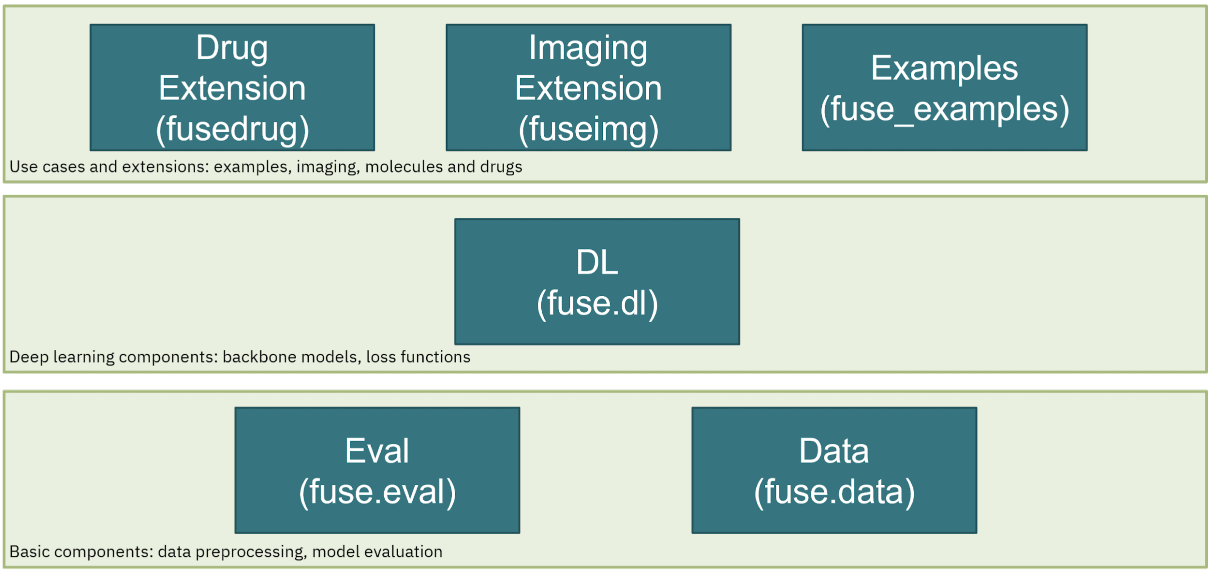
Drug Extension (fusedrug)
Imaging Extension (fuseimg)
Examples (fuse_examples)
Use cases and extensions: examples, imaging, molecules and drugs
DL (fuse.dl)
Deep learning components: backbone models, loss functions
Eval (fuse.eval)
Data (fuse.data)
Basic components: data preprocessing, model evaluation

Figure 6.1: BiomedSciAI FuseMedML conceptual architecture

Users and contributors seeking more information about FuseMedML can refer to the associated JOSS paper (Golts, 2023) and a blog post published in the PyTorch ecosystem (Raboh, 2022).

7. Glossary

AUC	Area Under the Curve
BP	Blood pressure
CCS	Clinical Classifications Software
DSS	Decision Support System
ECG	electrocardiography
EHR	Electronic Health Record
HCP	Healthcare Professional
HR	Heart rate
HRV	Heart Rate Variability
ICSM	Istituti Clinici Scientifici Maugeri hospital
MRI	Magnetic Resonance Imaging
NKI	Netherlands Cancer Institute
PPG	photoplethysmography
SPO2	Oxygen saturation
UI	User Interface
VOI	Volume of Interest

8. References

Barkan, E., Gilboa-Solomon, F., Quaglini, S. (2020): CAPABLE D5.1: Data Ready for Modelling and Reasoning Development. Zenodo. https://doi.org/10.5281/zenodo.4540570

Barkan, E., Porta, C., Rabinovici-Cohen, S., Tibollo, V., Quaglini, S., and Rizzo, M. (2023): Artificial intelligence-based prediction of overall survival in metastatic renal cell carcinoma. Journal of Frontiers in Oncology 16;13:1021684.

Breiman, L., (2001): Random forests. *Machine learning*, *45*(1), pp.5-32.

Chen, T., Guestrin, C., (2016): Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd international conference on knowledge discovery and data mining, Pages 785-794.

DeLong, E.R., DeLong, D.M. and Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, pp.837-845.

Dosovitskiy, A. et al. (2020): An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

Hanley, J.A., McNeil, B.J. (1982): The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), pp.29-36.

Gilboa-Solomon, F., Quaglini, S., Kogan, Alexandra, Glaser, S., Medlock, S., Barkan, E. and Lisowska, A. (2021): CAPABLE D5.2: Framework Defined (including Patients' Needs) Based on Available Data and Modeling Approaches, Vols. , Zenodo, available at: https://doi.org/10.5281/zenodo.5159285

Golts, A., Raboh, M., Shoshan, Y., Polaczek, S., Rabinovici-Cohen, S., Hexter, E. (2023): FuseMedML: a framework for accelerated discovery in machine learning based biomedicine. Journal of Open-Source Software. 8(81):4943.

Lippmann, R., 1994. Book Review: Neural Networks, A Comprehensive Foundation, by Simon Haykin. International Journal of Neural Systems, 5(04), pp.363-364.

Lundberg, S.M., Lee, S.-I. (2017): A Unified Approach to Interpreting Model Predictions. Adv. Neural Inf. Process. Syst. 30, 4765–4774

Langner, T., Ostling, O., et al. (2020): Kidney segmentation in neck-to-knee body mri of 40,000 uk biobank participants. Scientific reports, 10(1):20963.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika, 12(2), pp.153-157.

Moor, M., Banerjee, O., Abad, Z.S.H. et al. (2023): Foundation models for generalist medical artificial intelligence. Nature 616, 259–265.

Peleg, M., Lanzola, L., Panzarasa, S., Parimbelli, E., Polce, F., Quaglini, S., … Vicente, V. (2020): CAPABLE D2.1: Requirements Table and Use Case Description. Zenodo. https://doi.org/10.5281/zenodo.4540452.

Saly, D.L., Eswarappa, M.G., Street, S.E., Deshpande, P. (2021): Renal Cell Cancer and Chronic Kidney Disease. Advances in Chronic Kidney Disease, Volume 28, Issue 5, Pages 460-468.

Rabinovici-Cohen, S., Fernandez, X., Rabinovici-Cohen, S., Fernández, X. M., Grandal Rejo, B., Hexter, E., Hijano Cubelos, O., Pajula, J., Pölönen, H., Reyal, F., and Rosen-Zvi, M. (2-22): Multimodal Prediction of Five-Year Breast Cancer Recurrence in Women Who Receive

Neoadjuvant Chemotherapy. Cancers, Volume 14, Issue 16, 3848. https://doi.org/10.3390/cancers14163848.

Rabinovici-Cohen, S., Tlusty, T., Fernández, X. M., and Grandal Rejo, B. (2022): Early prediction of metastasis in women with locally advanced breast cancer. Proc. SPIE 12033 Medical Imaging. https://doi.org/10.1117/12.2613169

Raboh, M. (2022): FuseMedML blog - https://medium.com/pytorch/fusemedml-a-framework-accelerating-ai-based-discovery-and-code-reuse-in-the-biomedical-field-1ac874db3903

Tran, D. et al. (2018): A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 6450-6459.

9. Annexes

9.1. Annex 1 – Sensors data: time series features

The below figures plot the entire time series of watches 22, 27, and 28. We can see in the plot that there are periods when the watch stopped collecting data. This is probably when the watch was charged. We can see that many outlier values exist. We can also see that HR measurements based on ECG&PPG are scarce.

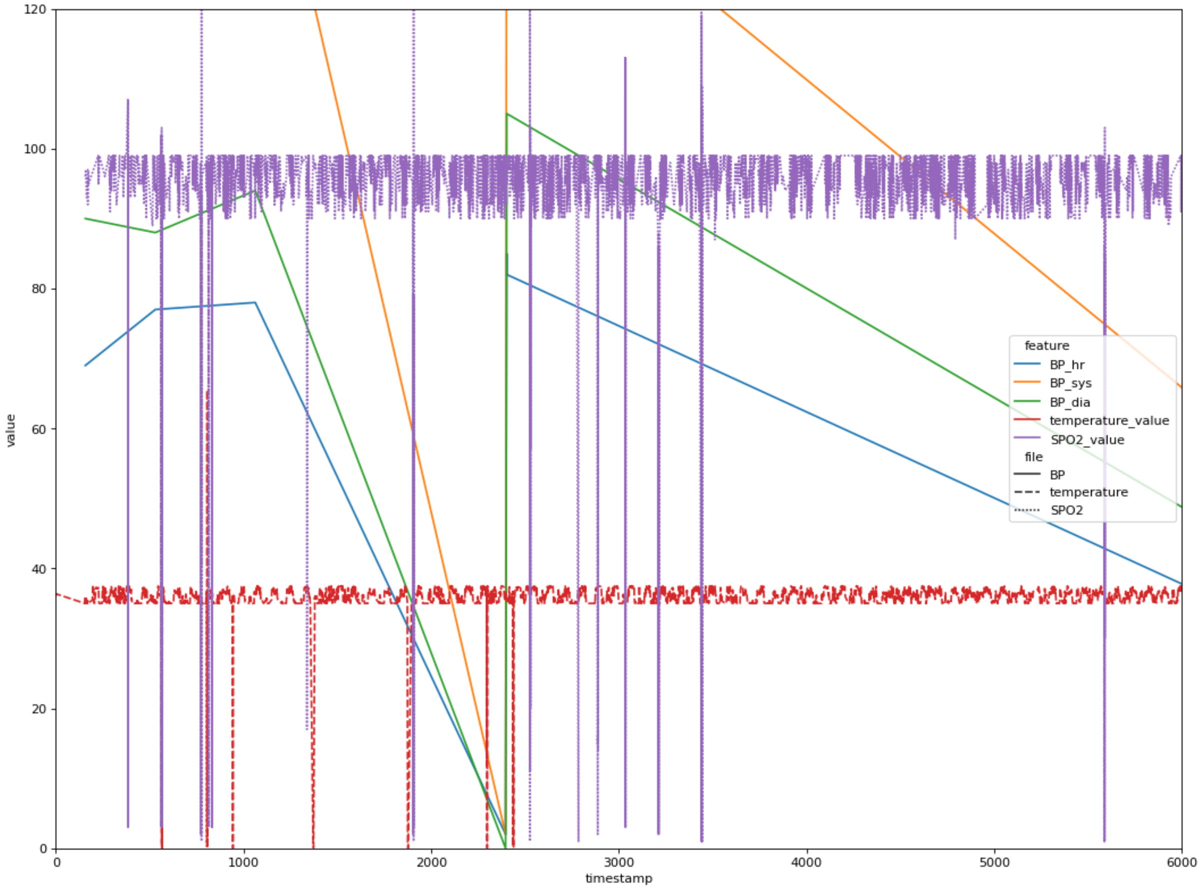

Figure 9.1.1: Watch 22 from ICSM pilot - all sensors time-series

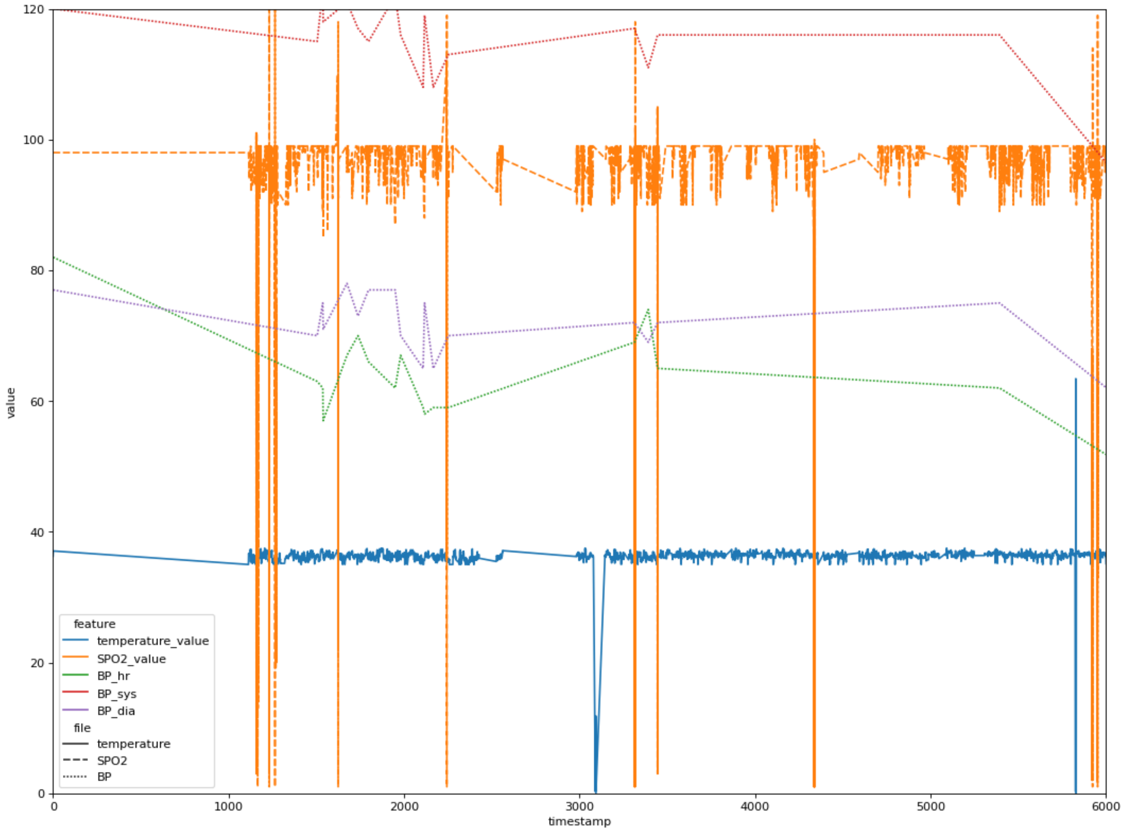

Figure 9.1.2: Watch 27 from ICSM pilot - all sensors time-series

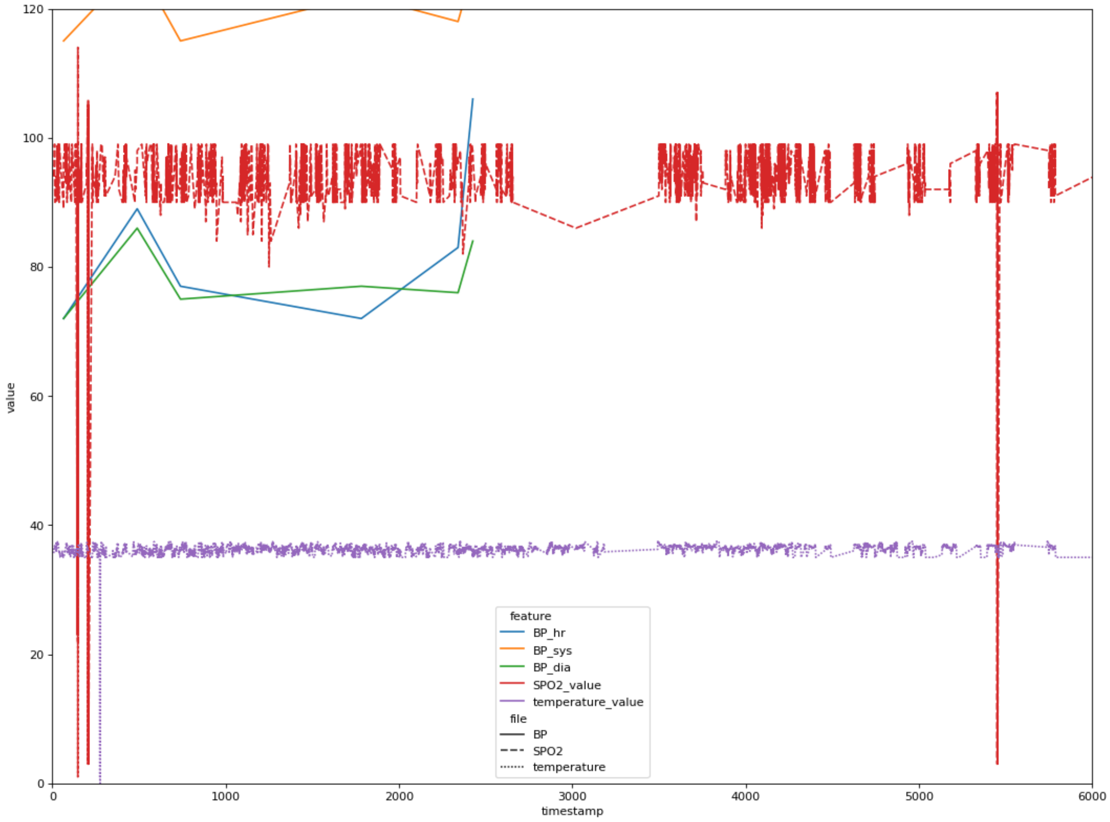

Figure 9.1.3: Watch 28 from ICSM pilot - all sensors time-series

9.2. Annex 2 – Sensors data: steps and calories

The total number of steps and calories during the period wearing the watch was calculated for the ICSM pilot as seen in the table below. We averaged over the total number of days wearing the watch, and added median heart rate, SPO2, systolic and diastolic blood pressure.

Table 9.2.1: ICSM pilot activity and median feature values

Watch	Days	Steps/day	Calories/day (cal/day)	BP_hr (bpm)	BP_sys (mmHg)	BP_dia (mmHg)	SPO2 percentage
VivoWatch 5-CAPABLE-0001	158.37	3718.07	185.90	61.5	115.5	74	96
VivoWatch 5-CAPABLE-0003	107.31	1859.41	92.92	88	97.5	56.5	97
VivoWatch 5-CAPABLE-0005	148.60	124.43	5.99	68	129	74	97
VivoWatch 5-CAPABLE-0006	180.87	1492.82	76.88	68	110	76	99
VivoWatch 5-CAPABLE-0009	20.07	3283.38	163.51	77	122	78	98
VivoWatch 5-CAPABLE-0011	127.35	6676.43	341.04	67	130	76	99
VivoWatch 5-CAPABLE-0014	38.51	1663.09	81.74	64	116	77	99
VivoWatch 5-	143.39	2164.14	108.54	78	123	79	97

CAPABLE-0017							
VivoWatch 5-CAPABLE-0020	154.07	1172.27	58.33	69	114	74	97
VivoWatch 5-CAPABLE-0021	204.28	3574.09	184.36	71	119	79	99
VivoWatch 5-CAPABLE-0022	181.66	4676.88	234.43	69	140	90	96
VivoWatch 5-CAPABLE-0025	148.23	246.77	12.26	83	114	82	98
VivoWatch 5-CAPABLE-0026	139.86	1647.76	105.44	70	124	81	
VivoWatch 5-CAPABLE-0027	154.36	2899.22	144.93	62	116	72	98
VivoWatch 5-CAPABLE-0028	122.44	1716.67	87.69	80	120	76.5	95
VivoWatch 5-CAPABLE-0029	56.62	174.92	8.60	70.5	120	70	95
VivoWatch 5-CAPABLE-0030	235.92	577.44	28.64	64	135	88	98

VivoWatch 5-CAPABLE-0031	147.82	318.35	15.88	78	120.5	82.5	99
VivoWatch 5-CAPABLE-0033	98.84	1639.48	117.69	86	123	74	94
VivoWatch 5-CAPABLE-0034	88.37	517.42	25.52	67	117	74	95
VivoWatch 5-CAPABLE-0035	77.13	1376.67	68.41	72	128	85	97
VivoWatch 5-CAPABLE-0036	119.97	5793.31	294.98	70	113	72	98
VivoWatch 5-CAPABLE-0041	3.31	3205.34	159.68				99
VivoWatch 5-CAPABLE-0046	48.67	3453.45	172.33	96	127	86	97
VivoWatch 5-CAPABLE-0047	56.55	3871.70	192.98				97
VivoWatch 5-CAPABLE-0049	38.67	3010.58	150.39	72.5	116	74.5	99
VivoWatch 5-CAPABLE-0053	9.79	2032.40	101.19				95

The total number of steps and calories during the period wearing the watch was calculated for the NKI pilot as well and can be seen in the table below. We averaged over the total number of days wearing the watch, and added median heart rate, SPO2, systolic and diastolic blood pressure.

Table 9.2.2: NKI pilot activity and median feature values

Watch	Days	Steps/day	Calories/day (cal/day)	BP_hr (bpm)	BP_sys (mmHg)	BP_dia (mmHg)	SPO2 percentage
VivoWatch5-CAPABLE-NKI-001	15.26	6754.59	326.98				98
VivoWatch5-CAPABLE-NKI-002	11.97	1726.17	86.28				98
VivoWatch5-CAPABLE-NKI-003	5.91	5442.36	271.20				99
VivoWatch5-CAPABLE-NKI-004	118.33	1287.97	64.67				99
VivoWatch5-CAPABLE-NKI-005	6.03	5228.95	262.04				93
VivoWatch5-CAPABLE-NKI-006	115.31	451.25	23.21				99
VivoWatch5-CAPABLE-NKI-007	168.88	3828.66	199.56	83	115.5	74.5	99
VivoWatch5-CAPABLE-NKI-008	24.65	964.55	48.49	64.5	130	82	96
VivoWatch5-CAPABLE-NKI-009	18.39	1198.75	59.86	91.5	134.5	85	99

VivoWatch5-CAPABLE-NKI-010	110.27	135.00	6.74	91	122.5	78.5	98
VivoWatch5-CAPABLE-NKI-011	29.93	1068.45	52.19				95
VivoWatch5-CAPABLE-NKI-012	46.28	3056.11	152.25				99
VivoWatch5-CAPABLE-NKI-013	5.85	243.73	11.97				94.5
VivoWatch5-CAPABLE-NKI-014	5.88	20628.96	1027.01				98
VivoWatch5-CAPABLE-NKI-016	7.68	8704.06	468.26				97
VivoWatch5-CAPABLE-NKI-017	94.48	2740.53	137.01				97
VivoWatch5-CAPABLE-NKI-019	3.24	2863.22	143.58				96
VivoWatch5-CAPABLE-NKI-020	1.76	2598.86	128.98				99
VivoWatch5-CAPABLE-NKI-023	27.89	705.43	85.92				98
VivoWatch5-CAPABLE-NKI-024	17.65	938.14	46.52				99
VivoWatch5-CAPABLE-NKI-026	33.63	1167.82	61.14	79	120	79	99

9.3. Annex 3 – Multimodal prediction models for task 1

Task 1 in our multimodal prediction models was to predict whether the patient with no CKD (CKD 0) will progress to severe CKD (CKD 3, 4, 5, dial) within five years. The figure below shows the result of the ensemble model for task 1. We see that the multimodal ensemble model that combines all the unimodal models is better than each unimodal model alone and achieves 0.806 AUC. These are similar results to what we have seen for task 2.

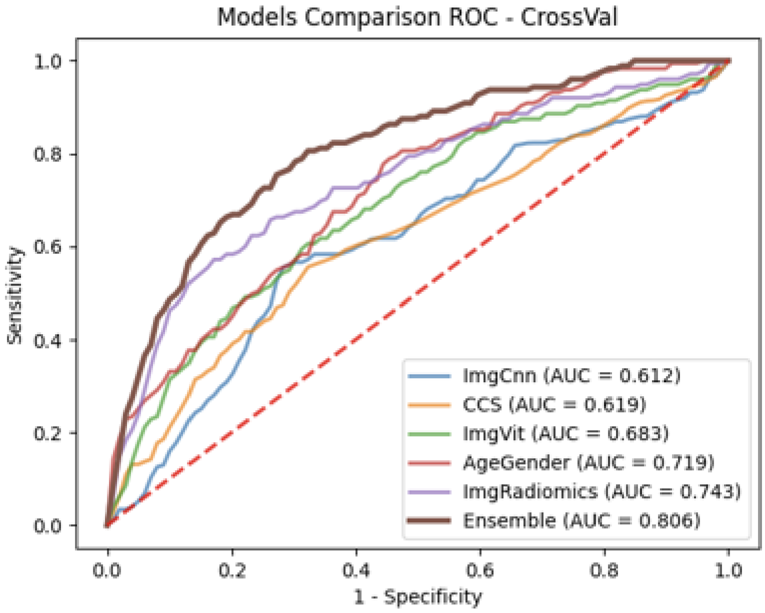

Figure 9.3.1: Ensemble model ROC curve for task 1

We also used the Delong test (Delong, 1988) to calculate the p-value when comparing the prediction of the individual models with those of the ensemble model. We received that the ensemble model is statistically better (p-value < 0.05) than each modality model alone. Once again these are similar results to what we got in task 2.

9.4. Annex 4 – Unimodal models results for task 2

The figure below shows the results of the demographics clinical model based on age and gender as well as its SHAP analysis. Logistic regression was found to be the best classifier for this kind of features. When averaging over the five cross-validation models, we get 0.704 AUC. The SHAP figure reveals that the age feature was much more important than the gender feature and that older male people tend to have higher risk for disease progression.

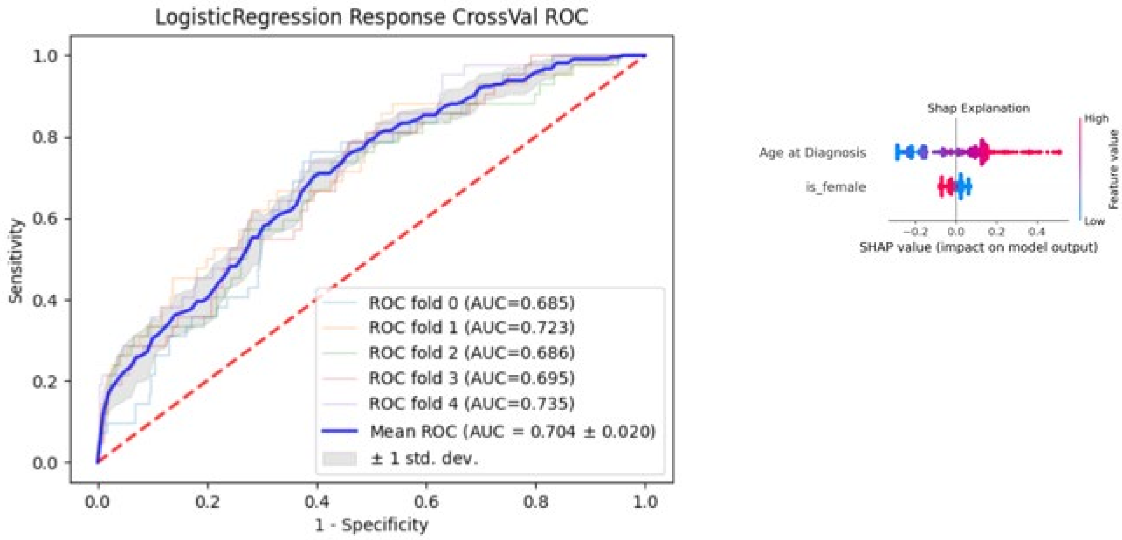

Figure 9.4.1: Demographics clinical model ROC curve and SHAP explanation

The figure below shows the results of the CCS clinical model based on CCS diagnosis codes. XGBoost was found to be the best classifier for this kind of features. When averaging over the five cross-validation models, we get 0.647 AUC.

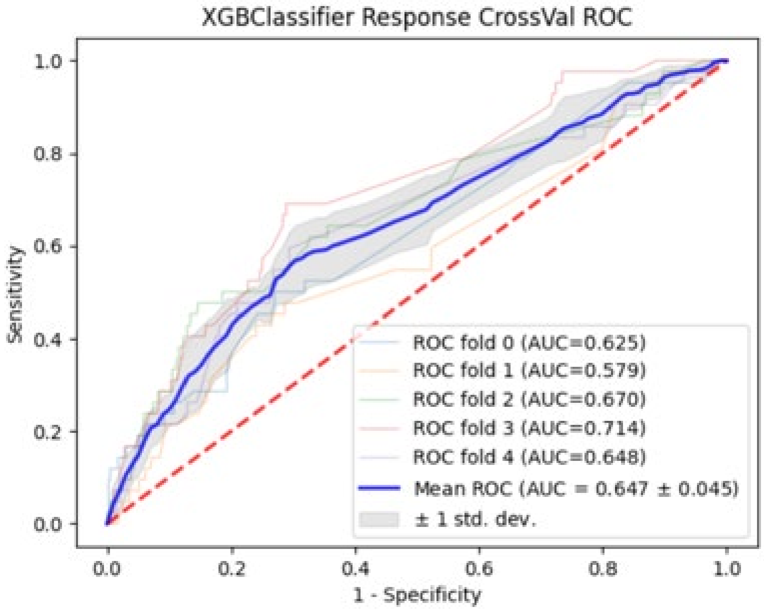

Figure 9.4.2: CCS clinical model ROC curve

The figure below shows the results of the CNN imaging model. Random forest was found to be the best classifier for this kind of features. When averaging over the five cross-validation models, we get 0.652 AUC. We also added interpretation of the model by using GradCam algorithm to find the area in the image that contributed the most to the prediction. We see that areas in the kidney and areas in the heart contributed the most to the prediction. This complies with what is known from clinical practice that heart disease is related to kidney disease.

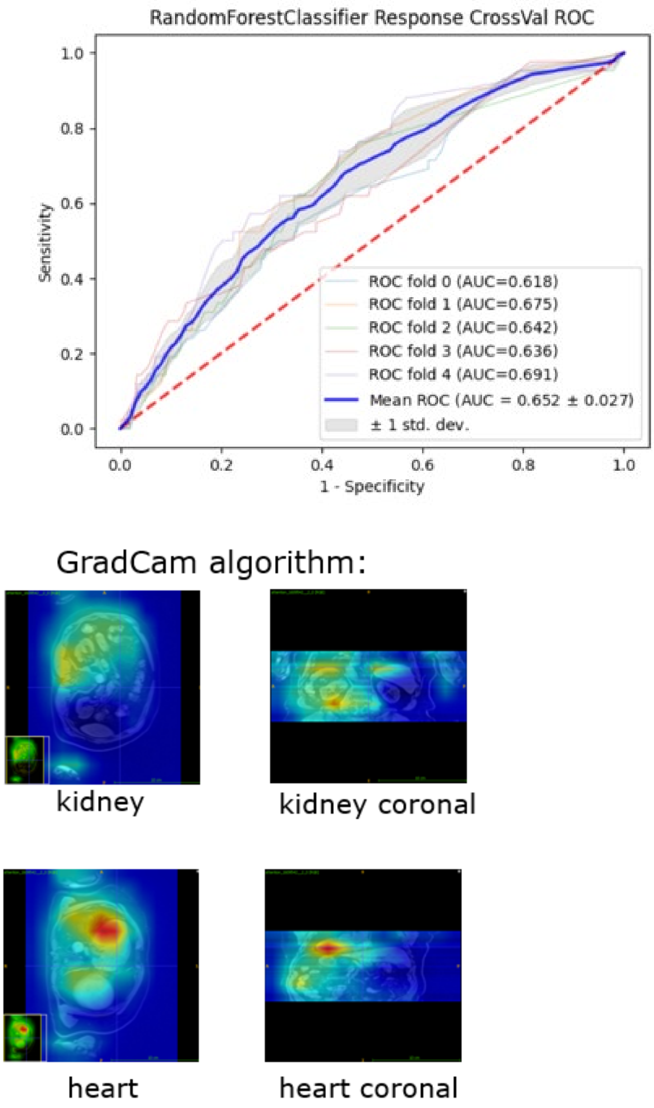

Figure 9.4.3: CNN imaging model ROC curve and GradCam explanation

The figure below shows the results of the ViT imaging model. Logistic regression was found to be the best classifier for this kind of features. When averaging over the five cross-validation models, we get 0.659 AUC.

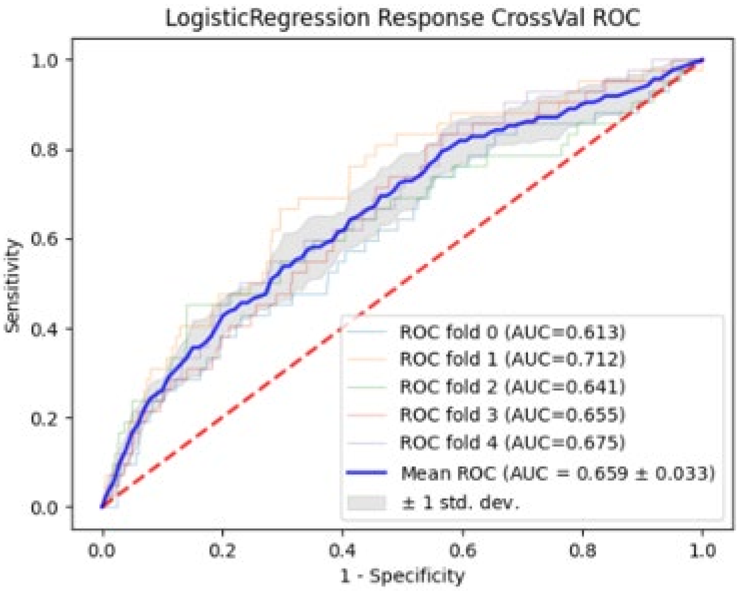

Figure 9.4.4: ViT imaging model ROC curve