PREDICTING PERIODS OF HYPERTENSION FOR PATIENTS USING REMOTE MONITORING DATA

A Thesis

by

JULIAN L. BEAULIEU

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Bobak J. Mortazavi
Committee Members,	Hye-Chung Kum
	Mark A. Lawley
Head of Department,	Scott Schaefer

December 2022

Major Subject: Computer Science

ABSTRACT

Hypertension is directly linked with an increase in mortality risks. This increase is especially observed in rural areas with uninsured or under-insured patients. Remote patient monitoring is a technological development that can help monitor hypertensive patients and detect changes within their blood pressure levels. The technological limitation is the inability to predict periods of hypertension. In this thesis, we address this issue by presenting a software system designed to successfully predict periods of hypertension and alert clinicians. Central to our software system is a framework for preparing data for predictive machine learning models, and an adaptive model for different sized sliding windows, to enable adverse health event predictions before the optimal input length has been reached. Using this framework, an XGBoost model only tuned for unbalanced data achieved an area under the receiver operating characteristic curve score of 0.77 and area under the precision recall curve score of 0.76. Feature importance plots demonstrate that our framework can extract the most impactful features, which are common across models. This demonstrates the ability to transform previously unsuitable data into data well-suited for periods of hypertension prediction and use it to alert clinicians of hypertensive periods to facilitate early interventions.

DEDICATION

I would like to dedicate this work to my mother, Petra.

Without her support and encouragements, all of this would have never been possible.

Danke, Mama.

ACKNOWLEDGMENTS

I would like to thank Julia, Marius, Felix, and Tom for always being there for me, celebrating my highs, and having my back through my lows.

I would also like to express great appreciation for Sorush for helping me figure out what and when I need to submit things to graduate on time. Michelle for supporting me with my research. I would also like to thank Dr. Bobak Mortazavi for letting me join the STMI lab and taking me in as one of his students. In the same breath, I would like to thank the STMI lab for helping me understand new concepts in machine learning and statistics.

Finally, I would like to thank the Texas A&M University Graduate and Professional School for an amazing LaTeX thesis template.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professor Mortazavi, and Professor Kum of the Department of Computer Science and Engineering, and Professor Lawley of the Department of Industrial Engineering. The work was also supported by Professor Erraguntla of the Department of Industrial and Systems Engineering

The data used was provided by Coordination Centric, and prepared by Sulki Park.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by the grant “Effective Real-World Telemonitoring of Chronic Disease for the Underserved” funded by the Texas AM President’s Office X-grant initiative.

NOMENCLATURE

SHAP	SHapley Additive exPlanations
ANN	Artificial Neural Network
LR	Linear Regression
RF	Random Forest
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic Curve
AUCPR	Area Under the Precision Recall Curve
BP	Blood Pressure
SBP	Systolic Blood Pressure
DBP	Diastolic Blood Pressure
GUI	Graphical User Interface
CVD	Cardiovascular Disease

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGMENTS iv

CONTRIBUTORS AND FUNDING SOURCES v

NOMENCLATURE vi

TABLE OF CONTENTS vii

LIST OF FIGURES ix

LIST OF TABLES x

1. INTRODUCTION 1

    1.1 Background 1
    1.2 Previous Work 1
    1.3 Objective 3

2. REMOTE MONITORING SYSTEM 4

    2.1 Data Preprocessing 4
        2.1.1 Dataset 5
        2.1.2 Multiple Readings 7
        2.1.3 Measurement Period and Missing Data Imputation 8
        2.1.4 Windowing 10
        2.1.5 Feature Extraction 11
        2.1.6 Labels and Event Horizon 14
        2.1.7 Feature Selection 14
    2.2 Event Prediction 15
        2.2.1 Classifier Choices 15
        2.2.2 Hyper Parameter Tuning 15
        2.2.3 Testing Data using Logistic Regression, Random Forest, and XGBoost 16
        2.2.4 Adaptive Model 16

3. PERFORMANCE EVALUATIONS 17

3.1 Introduction 17
3.2 Confusion Matrices 17
3.3 AUCROC 17
3.4 AUCPR 18
3.5 SHAP 18

4. RESULTS & DISCUSSION 19

4.1 Results 19
4.1.1 Hypertension Detection Performance 19
4.1.2 Model Overview 22
4.1.3 Adaptive Model Overview 27

5. CONCLUSION & FUTURE WORK 30

5.1 Discussion 30
5.2 Limitations and Future Directions 31
5.3 Conclusion 31

REFERENCES 33

LIST OF FIGURES

FIGURE Page

2.1 Architecture of proposed software system 5

2.2 Exclusion criteria flowchart for participants 6

2.3 Number of instances where a certain gap length occurred in the data. 9

2.4 7 Day Sliding Window with 7-Day Event Horizon 10

2.5 Adaptive Classification Model Architecture 16

4.1 Logistic Regression SHAP plot with 7-day event horizon and 3-day sliding window. 22

4.2 Logistic Regression SHAP plot with 7-day event horizon and 7-day sliding window. 23

4.3 Logistic Regression SHAP plot with 7-day event horizon and 8-day sliding window. 23

4.4 Logistic Regression SHAP plot with 7-day event horizon and 13-day sliding window. 24

4.5 Logistic Regression SHAP plot with 7-day event horizon and 14-day sliding window. 24

4.6 XG Boost SHAP plot with 7-day event horizon and 3-day sliding window. 25

4.7 XG Boost SHAP plot with 7-day event horizon and 7-day sliding window. 25

4.8 XG Boost SHAP plot with 7-day event horizon and 8-day sliding window. 26

4.9 XG Boost SHAP plot with 7-day event horizon and 13-day sliding window..... 26

4.10 XG Boost SHAP plot with 7-day event horizon and 14-day sliding window..... 27

4.11 AUCPR values for each sliding window submodel in adaptive model 28

4.12 AUCROC values for each sliding window submodel in adaptive model 29

LIST OF TABLES

TABLE Page

1.1 Blood pressure categories in mmHg. 1

2.1 Population characteristics of various minimum spell length requirements. 7

2.2 A table of features including their equation features was extracted instead of measured. 12

4.1 AUCROC results for each tested method across different sliding window, and event horizons sizes, with some results highlighted which are discussed in this section. 20

4.2 AUCPR results for each tested method across different sliding window, and event horizons sizes. 21

1. INTRODUCTION

1.1 Background

Elevated levels of blood pressure are becoming an increasingly large problem. In the United States alone, cardiovascular disease (CVD) is a leading cause of death and amounts to a significant part of the overall health care spending. With annual costs rising from $212 billion in 1996 to an estimated $320 billion in 2016, it is increasingly becoming a financial issue as well [1]. Fortunately, if CVD is detected early enough, treatment can reduce the risk of CVD-related deaths by 50-80% [2, 3, 4, 5, 6]. One treatable condition directly linked to CVD-related death is hypertension, which is indicated by sustained, increased levels of systolic (SBP) and diastolic blood pressure (DBP) as can be seen in 1.1 [7, 8].

Category	SBP in mmHg		DBP in mmHg
Normal	< 120	and	< 80
Elevated	120-129	and	< 80
Hypertension I	130-139	or	80-89
Hypertension II	≤ 140	or	≤ 90

Table 1.1: Blood pressure categories in mmHg.

Hypertension can be especially difficult to treat in populations that are uninsured, under-insured, rural or underrepresented [9]. Remote patient monitoring devices can help fill this gap by providing healthcare that is inexpensive and accessible. It provides a positive impact on provider and client satisfaction, as well as a decrease in the hospitalizations of clients in rural areas. With early information about hypertension, clinicians can act sooner to prevent hypertensive events [10].

1.2 Previous Work

An appropriate mechanism for remote patient monitoring and treatment of hypertension is necessary. Telemonitoring solutions to improve clinical care, particularly with a focus on cardiology,

have been studied extensively to only limited success, thus far.

Chaudhry et al. developed a telemonitoring system to predict and reduce readmission of patients diagnosed with heart failure [11]. In this study, patient-reported assessments of symptoms and general health led to alerts for call-center nurse interventions if daily values raised alarms. The study found no statistically significant improvement in readmission rates in the telemonitoring arm, despite modest baseline risk estimation of readmission risk [12]. Ong et al. conducted a similar remote patient monitoring study on heart failure patients, using more advanced remote sensing and risk estimating for intervention [13]. However, their study similarly failed to provide a statistically significant reduction in readmission. A subsequent study using the remote monitoring system designed for CVD patients did find improvement in predicting outcomes using baseline data and the first month of intervention data [14]. Beyond clinical trials, Park et al. analyzed real-world Medicaid telemonitoring data to characterize the adherence of patients with hypertension to telemonitoring with respect to blood pressure control. They found that there was a positive correlation associated with adherence and blood pressure control [15]. Abrar et al. designed a system with which blood pressure levels of the next 24h can be predicted by providing hourly blood pressure measurements through an app [16]. The app then sends the data to the cloud, at which it is prepared and evaluated by a combination of complex machine learning algorithms, before sending a forecast back to the app. In this system, the focus lies on a complex combination of genetic algorithms which try to establish a precise forecast given the past 24h of blood pressure data. While Abrar et al. used a 24-hour time period, and thus a 24 time step long sliding window, J. Lee et al. used a sliding window of 5.5 hours with time steps of 30 minutes [17] in order to try to predict upcoming hypertensive episodes. For this, they left a gap between their sliding window and the time period in which they were trying to predict a hypertensive event. They found that as the gap size increases, the ability to accurately predict an upcoming hypertensive event shrinks. Yet, J. Lee et al. were able to achieve an AUCROC score of 0.93.

1.3 Objective

The focus of this thesis will be the design of a remote monitoring system capable of reliably warning physicians of impending periods of hypertension. The proposed remote monitoring system will be able to automatically transmit blood pressure measurements from at-home blood pressure monitors to the cloud, prepare the data for a classifier, and then predict possible events. From here, physicians can intervene early and prevent the worsening of a patient's condition. The remote monitoring system has multiple sub systems to handle different tasks. The first task is to accept incoming data from the patient. A separate part of the system then transforms and prepares the incoming data for imputation. A machine learning model is then tasked with generating a prediction, which will be delivered to the physician. To better aid physicians, I also propose a model which adapts to the size of the input data size, thus enabling event predictions before the optimal sliding window size has been reached. Further, the data uploaded is used to continuously train and update the machine learning model used to generate the periods of hypertension prediction. A lot of focus is usually laid on developing a novel machine learning model for such a system. In this thesis, the focus lies on the preparing the data for a classifier. This allows for further development and improvement of classification accuracy in the future using more advanced models such as convolutional neural networks (CNN) or recurrent neural networks (RNN).

2. REMOTE MONITORING SYSTEM

2.1 Data Preprocessing

At the heart of the system is the framework which explores telemonitoring time-series data to identify key periods and features, leading to better machine learning models that effectively predict periods of hypertension, illustrated in Figure 2.1. This section describes the process by which the telemonitoring data is prepared for the classifiers and the architecture of the adaptive model. For the final models generated, the effectiveness of the predictions was computed using the area under the curve of the receiver operating characteristic (AUCROC) and area under the curve of precision and recall (AUCPR) to determine how best to separate those periods at risk from those not. The technique was then evaluated across linear and non-linear models that select predictive features (logistic regression with lasso regularization, random forest, and XGBoost).

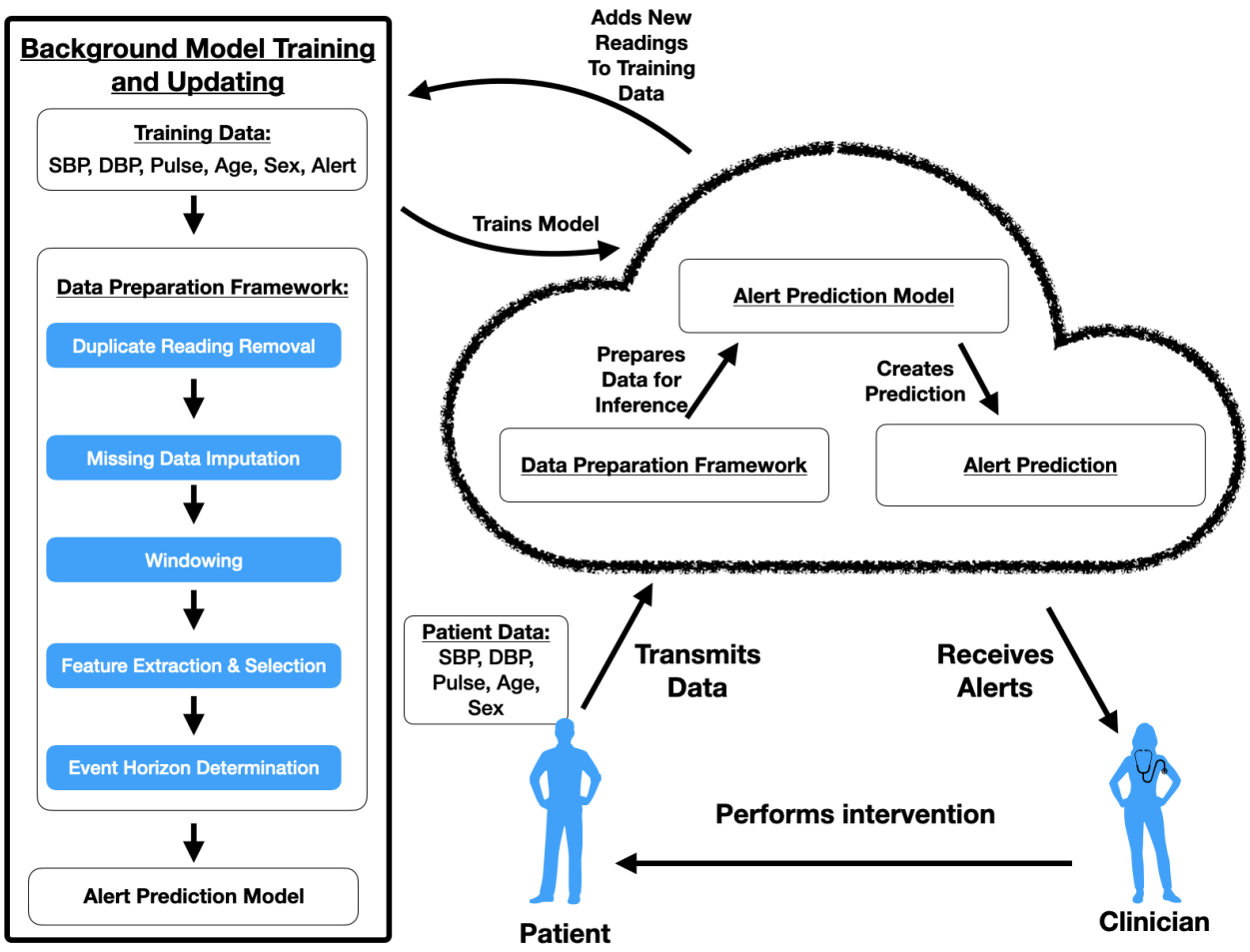

Figure 2.1: Architecture of proposed software system

2.1.1 Dataset

The data used to verify the effectiveness of my framework was provided by a telemonitoring company that uses data transmitted from Medicaid patients in Texas, USA. This data is transmitted from a Bluetooth-enabled blood pressure monitor via an intermediary device connected to the cloud. Three different device models were used in the study. The devices used were the Fora D40d, Taidoc 3223 and the A&D UA-767PBT. Roughly 1.6 million data points were gathered from 4,291 patients between August 2015 and January 2019. Of the 4,291 patients in the study, 1,494 (35%) were male and 2,797 (65%) female. The average age of the patients was 76.74 with a standard deviation of 13.29. Out of the total number of patients (N = 4,291), only between

2,053 and 2,906 patients were chosen due to several factors such as sparsity (lack of adherence) and duration (length of telemonitoring $<$ window size and event horizon) of data. Figure 2.2 visualizes the inclusion/exclusion criterion for patients. Table 2.1 shows the different population characteristics depending on how much data is used. This study was approved by the local IRB, Texas A&M Human Research Protection Program (IRB #IRB2018-0166D).

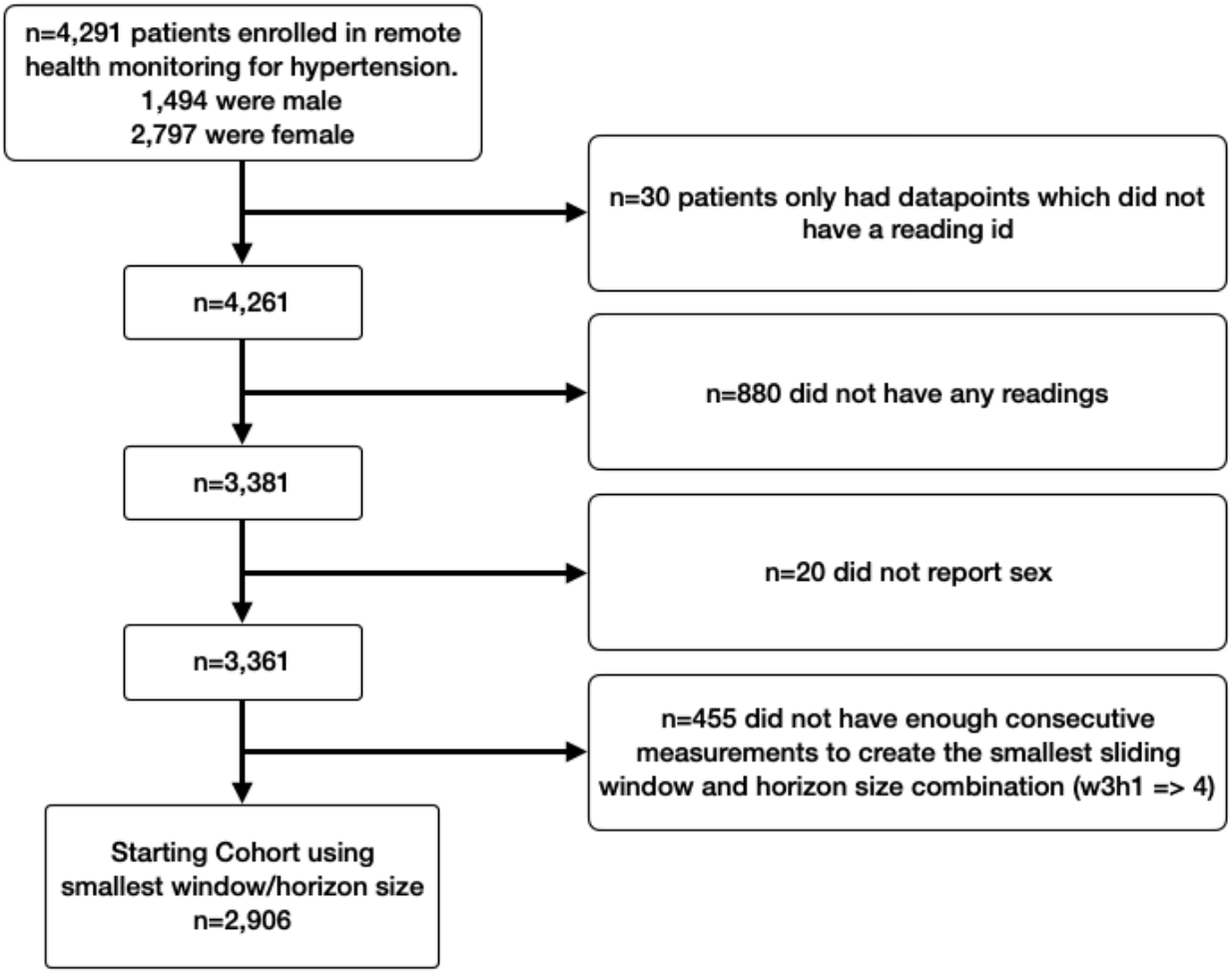

Figure 2.2: Exclusion criteria flowchart for participants

Population Characteristics Part 1

Minimum Data Length	n = 4	n = 5	n = 6	n = 7	n = 8	n = 9	n = 10	n = 11	n = 12	n = 13	n = 14	n = 15	n = 16
Number of PIDs:	2,906	2,854	2,795	2,745	2,685	2,645	2,611	2,561	2,533	2,491	2,451	2,413	2,379
Age Mean:	71.93	71.94	71.95	72.01	72.01	72.02	72.02	72.01	72.03	71.99	72.01	71.99	72.0
Age Std:	12.26	12.27	12.27	12.23	12.23	12.23	12.19	12.17	12.14	12.15	12.12	12.1	12.11
Sex Mean (M=0, F=1):	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
Spell Length Mean:	94.37	96.01	97.93	99.61	101.68	103.1	104.32	106.16	107.21	108.82	110.38	111.9	113.29
Spell Length Std:	115.11	115.5	115.95	116.33	116.78	117.09	117.35	117.74	117.97	118.3	118.62	118.93	119.21
Number of events Mean:	10.13	10.3	10.49	10.65	10.86	11.01	11.13	11.3	11.4	11.55	11.69	11.81	11.94
Number of events Std:	20.83	20.99	21.16	21.32	21.51	21.63	21.75	21.92	22.02	22.17	22.32	22.47	22.6
Number of imputed Measurements Mean:	8.63	8.77	8.94	9.08	9.26	9.36	9.46	9.6	9.69	9.82	9.94	10.06	10.16
Number of imputed Measurements Std:	12.16	12.22	12.29	12.36	12.44	12.5	12.55	12.62	12.67	12.73	12.8	12.86	12.92

Population Characteristics Part 2

Minimum Data Length	n = 17	n = 18	n = 19	n = 20	n = 21	n = 22	n = 23	n = 24	n = 25	n = 26	n = 27	n = 28
Number of PIDs:	2,348	2,315	2,280	2,245	2,217	2,193	2,167	2,141	2,123	2,097	2,078	2,053
Age Mean:	71.95	71.95	71.93	71.93	71.89	71.94	71.96	71.96	72.01	72.01	72.01	72.05
Age Std:	12.12	12.11	12.12	12.16	12.17	12.17	12.19	12.19	12.15	12.17	12.17	12.09
Sex Mean (M=0, F=1):	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.65	0.65	0.66
Spell Length Mean:	114.57	115.96	117.46	119.0	120.25	121.34	122.53	123.74	124.58	125.82	126.73	127.94
Spell Length Std:	119.46	119.74	120.03	120.33	120.57	120.77	121.0	121.23	121.4	121.64	121.82	122.05
Number of events Mean:	12.04	12.13	12.26	12.39	12.5	12.59	12.69	12.82	12.88	12.97	13.04	13.14
Number of events Std:	22.72	22.87	23.01	23.16	23.28	23.38	23.49	23.6	23.69	23.81	23.9	24.02
Number of imputed Measurements Mean:	10.25	10.35	10.47	10.59	10.69	10.77	10.85	10.94	11.01	11.09	11.16	11.24
Number of imputed Measurements Std:	12.97	13.03	13.1	13.16	13.21	13.26	13.31	13.36	13.39	13.44	13.48	13.54

Instead of showing population characteristic by window & horizon size, it is shown by length of data used. This is because a window of size three and horizon size two requires the same amount of data as a sliding window of size four and an event horizon of size one.

Table 2.1: Population characteristics of various minimum spell length requirements.

2.1.2 Multiple Readings

For each day the patient was in the study, the default protocol was one blood pressure reading per day. However, there were several situations when the patients would have duplicate readings such as due to faulty measurement devices, incorrect usage, or a reading outside a safe threshold prompting the nurse to check in and request another measurement. Using duplicate readings

could have a confounding effect by itself indicating intervention–i.e., rising blood pressure from repeated cuff inflation, or falling blood pressure due to nursing intervention. Thus, I chose the first measurement as the most representative of pre-intervention readings, and trusted that the machine learning models would appropriately handle the minimal noisy measurements.

2.1.3 Measurement Period and Missing Data Imputation

It is natural to find missing data on some days or some longer periods of missing data in real world data. However, using only the data we have, it is difficult to distinguish why there were spells of missing data followed by more monitoring data. Missing data could be due to a simple missed reading, or a more complex story, such as ending treatment on purpose (either by personal choice or clinician direction). In addition, in the real world, some patients may have multiple spells of treatment with varying degrees of gap between treatments. Thus, the more complicated situations beyond the first spell of monitoring has potential for more confounding effects from external factors that might bias the models. Because of the temporal nature of the data and the study aim to determine how many days of readings were necessary to appropriately generate hypertension alarms, I wanted to capture a relatively complete period of data. Too large a gap would render predictions potentially useless, while too small would inhibit realistic patient data modeling. To determine the optimal amount of imputation needed for measurement gaps and an appropriate end point for the first monitoring spell, I first analyzed the amount of each gap length. Figure 2.3 visualizes the amount of a certain gap length. For example, there were 48,549 instances of a patient missing a reading. This figure indicated that three days of imputation would be a good medium between retaining time-series accuracy and having a sufficient amount of data to properly train the model. Thus, patients missing more than three consecutive days (>3) of readings were assumed to have ended their treatment. Based on this definition, only the first spell of monitoring was used, on which imputation was done using forward fill on the SBP, DBP, and pulse data of each patient.

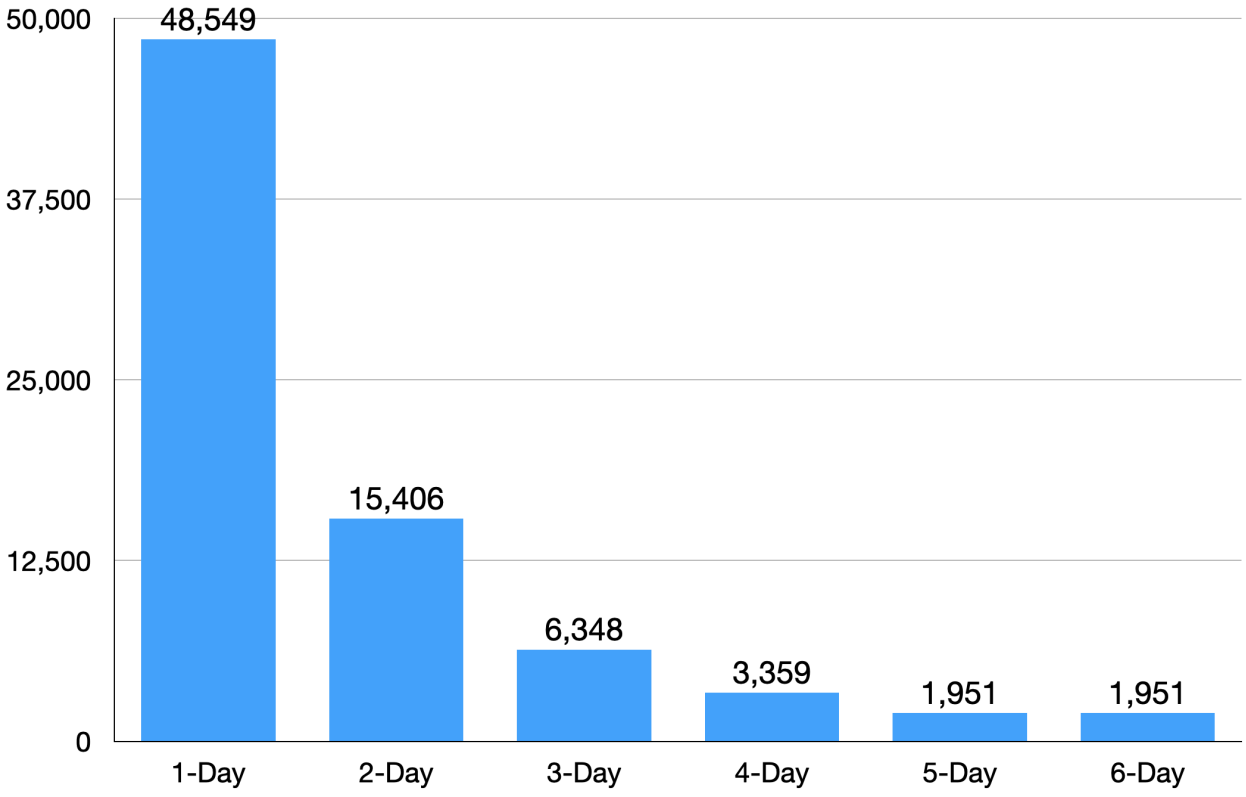

Figure 2.3: Number of instances where a certain gap length occurred in the data.

2.1.4 Windowing

I used a sliding window approach to train and test models for periods of hypertension. Figure 2.4 illustrates a seven-day sliding window, with an example of estimating periods of hypertension over the next 7 days (event-horizon prediction length discussed below). Larger sliding windows and/or event horizons need more data and thus fewer windows can be created, but they can potentially enable better modeling of blood pressure trends. Therefore, larger than seven days may eliminate early warnings, and requiring long periods of readings before the system may generate alarms. Alternatively, small window sizes may not have sufficient signals to generate accurate enough alarms. To research this effect, sliding window sizes between 3 and 14 days were used.

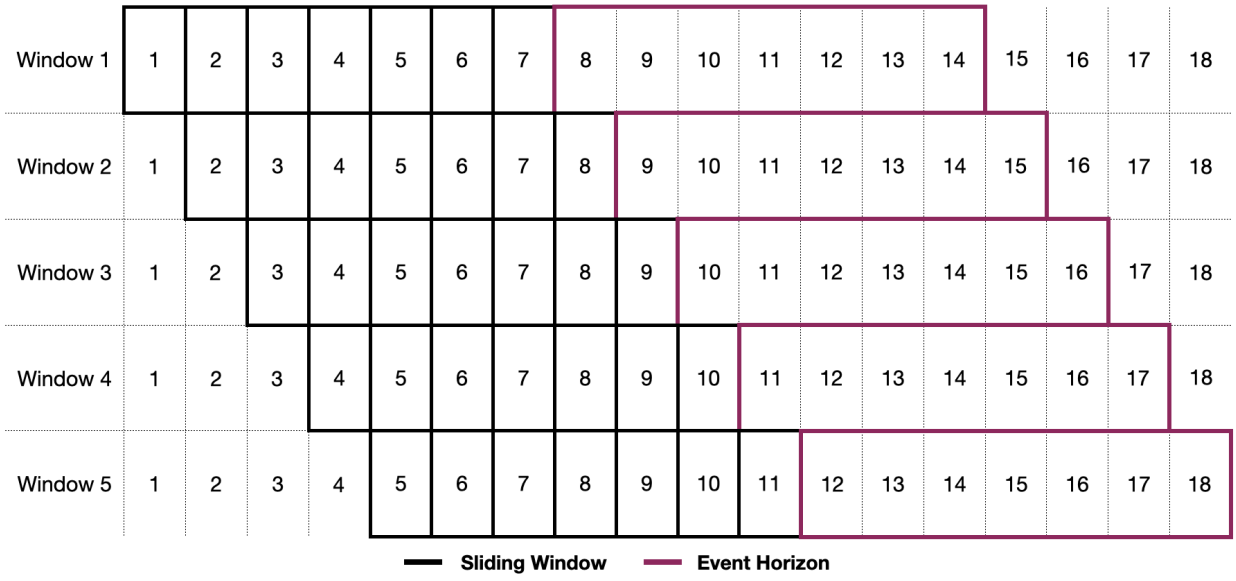

Figure 2.4: 7 Day Sliding Window with 7-Day Event Horizon

2.1.5 Feature Extraction

From the window of data that consisted of multiple SBP, DBP, and pulse readings, as well as additional data on age and sex of the patient, I extracted features to represent the trends in blood pressure measurements. Following similar work in the field of glucose monitoring and hyperglycemic excursions (periods of high blood sugar), I extracted the variance, slope, and maximum positive and negative changes of consecutive days [18]. Additionally, I extracted the difference between the first and last readings, and middle and last readings of the SBP, DBP, and pulse data points. These extracted features were chosen based on their ability to provide additional information on the variability of the blood pressure readings and trends towards the latter portion of the window. For instance, a high seven-day standard deviation means that the blood pressure of a patient fluctuated a lot, while a low standard deviation means that the blood pressure was consistent over the period. I then concatenated this into one input vector for each day. For example, if considering a seven-day sliding window, then I would have features 1-7 as raw SBP readings, 8-14 as raw DBP readings, 15-21 as raw pulse readings, then the next 18 features (22-39) would be the extracted ones from the sliding window and the final two features (40-41) would be age and sex, thus resulting in 41 features before moving on to feature elimination.

Feature List		
Feature Name	Type of Feature	Feature Generation
SBP Day 1 – w	Measured	N/A
DBP Day 1 – w	Measured	N/A
Pulse Day 1 – w	Measured	N/A
Slope of SBP/DBP/Pulse	Extracted	Slope of linear regression model
Standard Deviation of SBP/DBP/Pulse	Extracted	$\sqrt{\frac{\sum_{i=1}^{w}(x_i - \mu)^2}{w}}$
Max positive change between two days of SBP/DBP/Pulse	Extracted	See Algorithm 1
Max negative change between two days of SBP/DBP/Pulse	Extracted	See Algorithm 2
Difference between first & last day of SBP/DBP/Pulse window	Extracted	$x_w - x_1$
Difference between first & middle day of SBP/DBP/Pulse window	Extracted	$x_{w/2} - x_1$
Age	Measured	N/A
Sex	Measured	N/A

'w' stands for the length of a window.

Table 2.2: A table of features including their equation features was extracted instead of measured.

Algorithm 1 Algorithm to calculate the max positive change between two adjacent days of SBP, DBP, or Pulse.

```
procedure MAXPOSITIVECHANGE(w)
    max_change ← inf
    for i ← 1 to length(w) do
        if (w[i-1] - w[i]) > max_change then
            max_change ← w[i-1] – w[i]
        end if
    end for
    return max_change
end procedure
```

Algorithm 2 Algorithm to calculate the max negative change between two adjacent days of SBP, DBP, or Pulse.

```
procedure MAXNEGATIVECHANGE(w)
    max_change ← -inf
    for i ← 1 to length(w) do
        if (w[i-1] - w[i]) < max_change then
            max_change ← w[i-1] – w[i]
        end if
    end for
    return max_change
end procedure
```

2.1.6 Labels and Event Horizon

Hypertensive events were defined as a set of measurements which were above a clinician specified threshold for SBP (personal thresholds set by clinicians). Then, an alert label indicated the presence of a hypertensive event during the length of the event horizon, where "1" denoted the presence of a hypertensive event and a "0" denoted no hypertensive event. The length of the event horizon is important in determining how early a prediction can be made. A small event horizon, for example, could potentially be more precise but may not provide sufficient time for intervention, and be harder to predict. Conversely, a large event horizon could potentially provide sufficient time for intervention, and a potential higher prediction accuracy, (i.e., whether an event occurred in a given time period) but result in reduced precision. The data preparation step tries to best identify the optimal prediction event horizon, and sliding window size. I tested event horizon sizes between 1 ad 14 days to identify accuracy in the data, and whether additional time would provide clinicians with an opportunity to intervene. Longer horizons, such as between 7 and 14 days, were tested to determine the impact longer horizons had on detection accuracy.

2.1.7 Feature Selection

While the machine learning models selected are capable of selecting features, the non-linear, higher-dimensional models (Random Forest and XGBoost) may select co-linear features in subsequent trees, which may result in a perceived reduced feature importance when providing clinicians with an estimation of hypertension and what is leading to that estimation (to help guide interventions). Therefore, after training, I applied backward feature elimination using the validation set and logistic regression to remove co-linear features. Logistic regression's L1 regularization would remove unnecessary features, and help select the key features that are directly related, with statistical test and p-value illustrating this relationship. Using a backwards elimination based upon the p-values, (removal of the highest p-value), I ultimately removed all features with a p-value greater than 0.05. I then used this feature subset, without co-linear features, to train Random Forest and XGBoost.

Each model uses a proportion of the features generated based upon window size and horizon length. However, while the number of features select varies (from min of x proportion of features used to max of y proportion of generated features used), there is consistency across feature importance. For one, all SBP measurements, age, and sex are kept during feature selection. Additionally, the standard deviation, and slope of SBP along with DBP, and Pulse data are often kept. The max positive/negative change along with the difference between first and last/middle features are often chosen as well, although it differs whether the SBP, DBP, or pulse version are kept. These are evident across Figures 4.5 to 4.10.

2.2 Event Prediction

2.2.1 Classifier Choices

Although there is a wide range of classifiers that would have also been well-suited for this task, I decided on a logistic regression classifier, decision forest, and a XGBoost classifier. With these methods, it is easy to determine which features emerged as more important. This facilitated more tuning of the data preparation and cleaning step, as it allowed me to pinpoint which extracted features worked well and which ones did not. Using the logistic regression model's coefficients, I was able to determine what the classifier was using to make its prediction. In the case of the decision forest, and XGBoost I was able to look at the individual decision trees for information on what was most impactful for their classification.

2.2.2 Hyper Parameter Tuning

For logistic regression with L1 regularization, I used the lambda regularization parameter. For XGBoost, training the depth of each tree was set to a max depth of two, the number of trees to 21, and the max step size remained unchanged. Additionally, I provided the models with a class-weight parameter because of the data set imbalance. For most window and event horizon sizes, there were more "0s" (no hypertensive events) than "1s" (hypertensive events) which resulted in an unbalanced data set. I computed the weight to be $= \frac{no\ event\ happened}{event\ happened}$ for the event happened class, to increase its importance to a "balanced" level.

2.2.3 Testing Data using Logistic Regression, Random Forest, and XGBoost

All classifiers were trained on the same set of training data. The data was split between training, validation, and testing with a ratio of 70:15:15 respectively based upon Patient IDs. The data was randomly shuffled at the patient level before splitting. The shuffling was done on the patient IDs rather than the entire data set to maintain continuity of the sliding windows, resulting in a patient being in either the training, validation, or testing data set, and demonstrating the model's ability to generalize to new, unseen patients.

2.2.4 Adaptive Model

To facilitate the possibility of earlier interventions and treatment by physicians, a model which can adapt to the varying sliding window input sizes is required. Traditional classifiers work with a fixed input size, thus preventing early predictions, such as using only 3 days of input data. The solution to this inability is to conglomerate the best classifiers for each input size to one larger model. As depicted in Figure 2.5, the input from the patient is first augmented using the extracted features discussed above, then passed into the model along with the length of the input sequence. From here, the adaptive model uses one of its sub models to create an event prediction, which ultimately gets sent to the physician. The benefit of the model design is the ability to continuously update and improve the sub-models, which in turn leads to a higher overall accuracy.

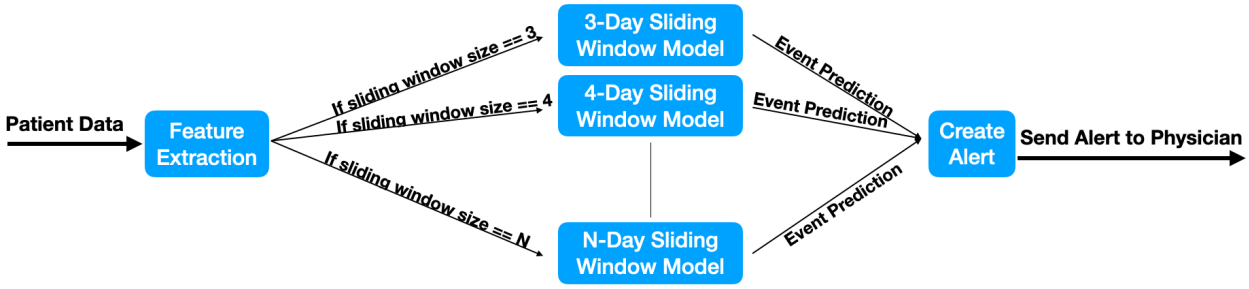

Figure 2.5: Adaptive Classification Model Architecture

3. PERFORMANCE EVALUATIONS

3.1 Introduction

The data used to train the machine learning classifiers were unbalanced, thus I decided not to use a performance measure such as Root-Mean-Square Deviation (RMSE) to determine the classification accuracy. To best determine the performance of the machine learning models used, I measure their performances using the area under the receiver operating characteristic curve (AUCROC), the area under the precision recall curve (AUCPR), confusion matrices, and SHapley Additive exPlanations (SHAP).

3.2 Confusion Matrices

Both AUCROC and AUCPR are derived from confusion matrices by indicating how many classification instances were classified as correct and incorrect. Further, they indicate if the correctly classified instances were of the true positive (TP), false positive (FP), true negative (TN), and false negative (FN) kind. These values give a good overview on how a classifier is performing. As previously stated, using a 50% risk probability as the threshold for predicting positive versus negative alerts may not yield the best results. Some classifiers perform better or worse depending on where the classification threshold is set. This means that there could be a multitude of different performances depending on the classification threshold set, and thus there can be a multitude of different confusion matrices.

3.3 AUCROC

Having a multitude of different confusion matrices that all show the binary classification performance of different thresholds is not very efficient. Thus, a metric such as AUCROC can indicate how the model is performing using different thresholds. AUC stands for area under the curve and is primarily useful for comparing different ROC plots with each other to determine the best classifier. The ROC, or receiver operating characteristic, takes the true positive and false positive rate of different binary classification thresholds and plots them against each other. This allows us to clearly

see which threshold yields the best classifications. Thus, using the AUCROC curve is particularly useful for seeing how well a classifier is able to discern between two classes.

3.4 AUCPR

Despite the effectiveness of using the AUCROC curve, Area Under the Precision Recall Curve (AUCPR) still adds valuable information, especially in cases of an imbalanced dataset. Imagine having a dataset which is highly unbalanced–e.g., 90% of class 0 and 10% of class 1, where class 1 represents cancer found in a patient. A regular machine learning algorithm might find it hard to predict cancer instances and would predict 0 most of the time. That could lead to a high accuracy of let's say 0.95 and an AUCROC score of 0.93, but with the obvious oversight that the data is highly imbalanced, and so predicting case 0 or no cancer is not as valuable as predicting cancer. This is why we also use AUCPR. The area under the curve aspect stays the same – to compare different models with each other. However, the false positive rate is swapped with the precision ($Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$). This means we are focusing on the true positive, false positive, and false negative values and ignoring the true negative values. Thus focusing on how well we predict the positive case instead of the negative case. In the case of this research, predicting an upcoming period of hypertension is much more important than predicting the opposite. As such, using AUCPR is a key element of comparing performances between the different classifiers trained and used in this research.

3.5 SHAP

With the introduction of Lundberg et al.'s SHapley Additive exPlanations, there now is a better way to determine the importance of a feature to the classification model. Additionally, SHAP enables us to compare the classifiers with each other in terms of the feature importance. This is specifically important because it facilitates easily understanding what each model deems to be an important feature and what not.

4. RESULTS & DISCUSSION

4.1 Results

4.1.1 Hypertension Detection Performance

Table 4.2 presents the AUCPR values and Table 4.1 AUCROC values for various event horizon, and sliding window lengths. In general, as the event horizon length increases, performance improves. This is both a function of the amount of predictive data used to make an estimation and the increasing probability of a hypertensive event occurring as the horizon length increases. While the AUCPR performance of the classifiers improved with an increase in both sliding window size and event horizon size, the same cannot be said about AUCROC. Here we can see that the XG Boost classifiers do not see a substantial performance gain as the sliding window or horizon size increase (colored rows). Further, we can notice that the AUCROC values start to converge at around 8-day of sliding window length and a 14-day event horizon (bold numbers). This indicates that the increasing window lengths do not make the problem easier to model, which could be a result from the reduction in patients seen in Table 2.1.

		Event Horizon Size												
Sliding Window Size		**1 Day**	**2 Day**	**3 Day**	**4 Day**	**5 Day**	**6 Day**	**7 Day**	**8 Day**	**9 Day**	**10 Day**	**11 Day**	**12 Day**	**13 Day**	**14 Day**
3 Day	**Logistic Regression**	0.58	0.63	0.66	0.66	0.68	0.70	0.70	0.72	0.71	0.74	0.74	0.74	0.73	0.72
	Random Forest	0.57	0.62	0.66	0.65	0.68	0.69	0.70	0.71	0.71	0.74	0.73	0.73	0.72	0.72
	XG Boost	0.78	0.76	0.76	0.74	0.73	0.75	0.75	0.75	0.74	0.76	0.74	0.74	0.73	0.73
7 Day	**Logistic Regression**	0.60	0.63	0.65	0.70	0.71	0.74	0.74	0.74	0.76	0.75	0.77	0.76	0.76	0.75
	Random Forest	0.57	0.62	0.64	0.68	0.70	0.73	0.72	0.74	0.75	0.74	0.76	0.75	0.75	0.75
	XG Boost	0.80	0.78	0.75	0.77	0.77	0.78	0.76	0.77	0.78	0.77	0.77	0.77	0.76	0.76
8 Day	**Logistic Regression**	0.61	0.66	0.68	0.69	0.72	0.74	0.75	0.75	0.77	0.76	0.78	0.75	0.76	**0.80**
	Random Forest	0.60	0.64	0.67	0.68	0.72	0.73	0.74	0.75	0.76	0.76	0.77	0.74	0.76	**0.79**
	XG Boost	0.79	0.78	0.76	0.76	0.78	0.79	0.78	0.78	0.77	0.78	0.79	0.75	0.78	**0.80**
13 Day	**Logistic Regression**	0.59	0.63	0.72	0.70	0.74	0.74	0.74	0.76	0.77	0.79	0.76	0.78	0.79	**0.80**
	Forest	0.57	0.62	0.70	0.69	0.73	0.72	0.73	0.75	0.76	0.78	0.76	0.77	0.78	**0.80**
	XG Boost	0.80	0.78	0.80	0.78	0.79	0.80	0.77	0.79	0.79	0.80	0.77	0.79	0.80	**0.81**
14 Day	**Logistic Regression**	0.61	0.64	0.68	0.70	0.72	0.71	0.76	0.77	0.78	0.77	0.76	0.76	0.77	0.75
	Random Forest	0.59	0.64	0.66	0.68	0.71	0.70	0.75	0.76	0.77	0.76	0.75	0.76	0.77	0.75
	XG Boost	0.79	0.79	0.77	0.78	0.77	0.77	0.79	0.80	0.78	0.78	0.76	0.77	0.78	0.77

Table 4.1: AUCROC results for each tested method across different sliding window, and event horizons sizes, with some results highlighted which are discussed in this section.

		Event Horizon Size												
Sliding Window Size		**1 Day**	**2 Day**	**3 Day**	**4 Day**	**5 Day**	**6 Day**	**7 Day**	**8 Day**	**9 Day**	**10 Day**	**11 Day**	**12 Day**	**13 Day**	**14 Day**
3 Day	**Logistic Regression**	0.42	0.51	0.60	0.55	0.63	0.69	0.68	0.71	0.71	0.75	0.76	0.78	0.75	0.75
	Random Forest	0.40	0.48	0.59	0.55	0.62	0.67	0.67	0.70	0.71	0.74	0.75	0.78	0.74	0.75
	XG Boost	0.43	0.52	0.61	0.57	0.64	0.69	0.69	0.71	0.72	0.76	0.77	0.79	0.75	0.76
7 Day	**Logistic Regression**	0.45	0.54	0.56	0.66	0.66	0.73	0.72	0.75	0.77	0.74	0.79	0.78	0.76	0.80
	Random Forest	0.43	0.52	0.56	0.65	0.67	0.73	0.72	0.75	0.77	0.73	0.80	0.78	0.76	0.80
	XG Boost	0.47	0.54	0.57	0.67	0.68	0.74	0.73	0.76	0.77	0.74	0.80	0.79	0.77	0.80
8 Day	**Logistic Regression**	0.49	0.57	0.62	0.64	0.71	0.73	0.74	0.78	0.80	0.77	0.82	0.76	0.80	0.84
	Random Forest	0.49	0.57	0.62	0.63	0.70	0.73	0.74	0.77	0.79	0.76	0.81	0.75	0.80	0.84
	XG Boost	0.51	0.58	0.63	0.64	0.71	0.73	0.75	0.78	0.80	0.77	0.82	0.76	0.81	0.84
13 Day	**Logistic Regression**	0.45	0.53	0.70	0.64	0.72	0.71	0.72	0.77	0.80	0.80	0.80	0.80	0.81	0.85
	Random Forest	0.44	0.51	0.70	0.65	0.72	0.70	0.71	0.76	0.80	0.81	0.80	0.80	0.80	0.85
	XG Boost	0.46	0.53	0.70	0.66	0.73	0.71	0.72	0.77	0.81	0.81	0.80	0.80	0.81	0.85
14 Day	**Logistic Regression**	0.47	0.58	0.62	0.62	0.69	0.69	0.78	0.78	0.79	0.80	0.79	0.80	0.79	0.79
	Random Forest	0.46	0.57	0.61	0.62	0.67	0.69	0.77	0.78	0.78	0.79	0.78	0.80	0.78	0.79
	XG Boost	0.48	0.59	0.62	0.63	0.68	0.69	0.78	0.78	0.79	0.80	0.78	0.81	0.79	0.80

Table 4.2: AUCPR results for each tested method across different sliding window, and event horizons sizes.

4.1.2 Model Overview

I used SHAP [19] on both the Logistic Regression and XGBoost model to understand which features impacted the models most, illustrated in Figures 4.1 to 4.10. Using SHAP enables us to properly visualize the output of our machine learning model based on the importance of the input features and the range of those continuous features and their increase or decrease in risk. A higher SHAP value is indicated by the color red, whereas blue denotes a lower SHAP value. One can observe a clear emphasis on the importance of SBP readings over DBP and Pulse readings. These results are expected knowing that the main focus is predicting periods of hypertensive events based on SBP thresholds. Looking at the SHAP plots for 3, 7, 8, 10, 13, and 14 day sliding windows, we can see that typically the last 3-4 days of SBP values are the most important to the classifiers.

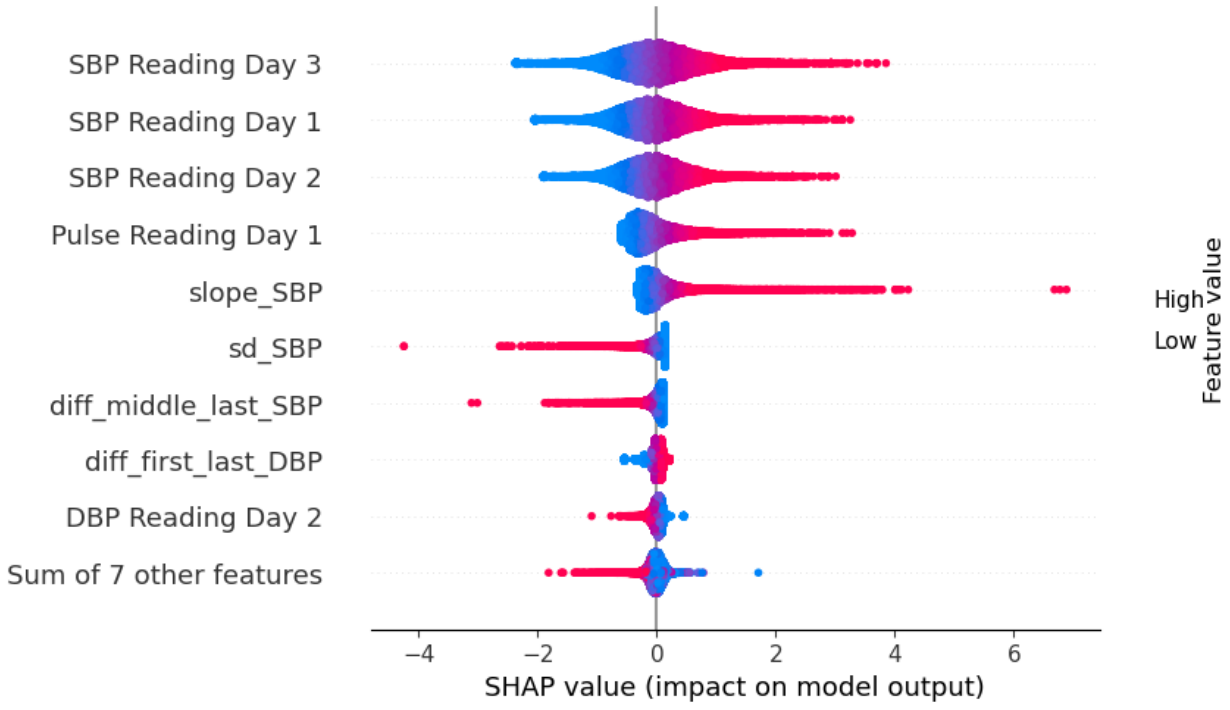

Figure 4.1: Logistic Regression SHAP plot with 7-day event horizon and 3-day sliding window.

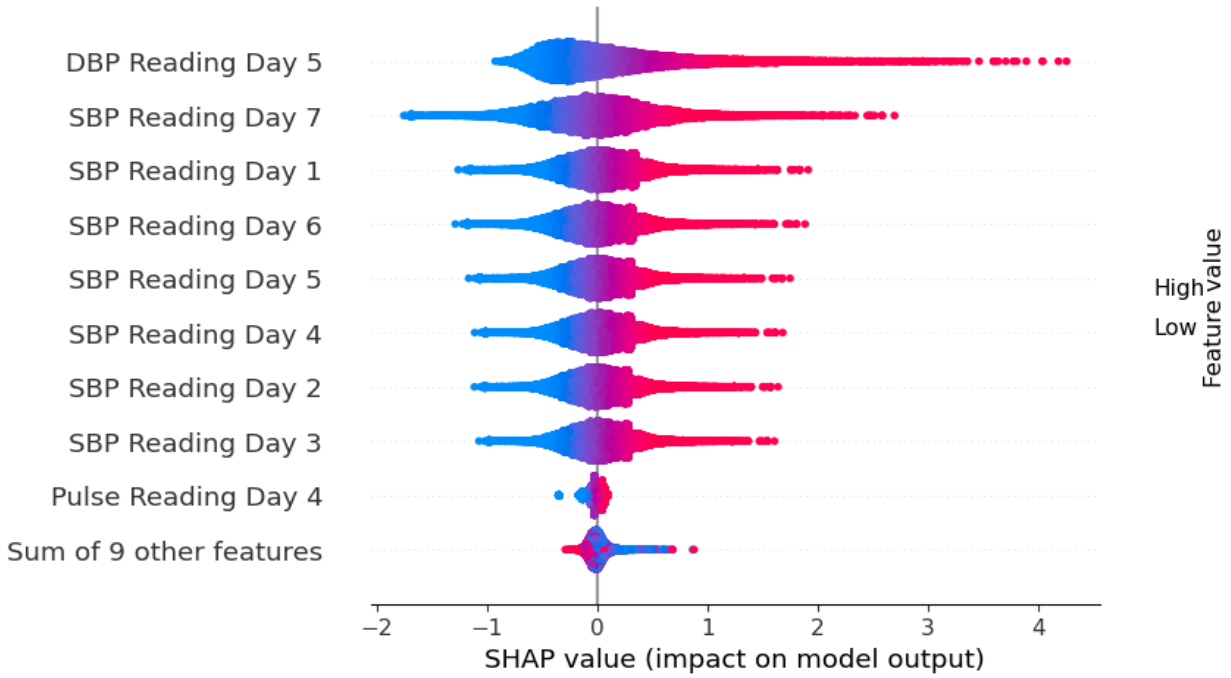

Figure 4.2: Logistic Regression SHAP plot with 7-day event horizon and 7-day sliding window.

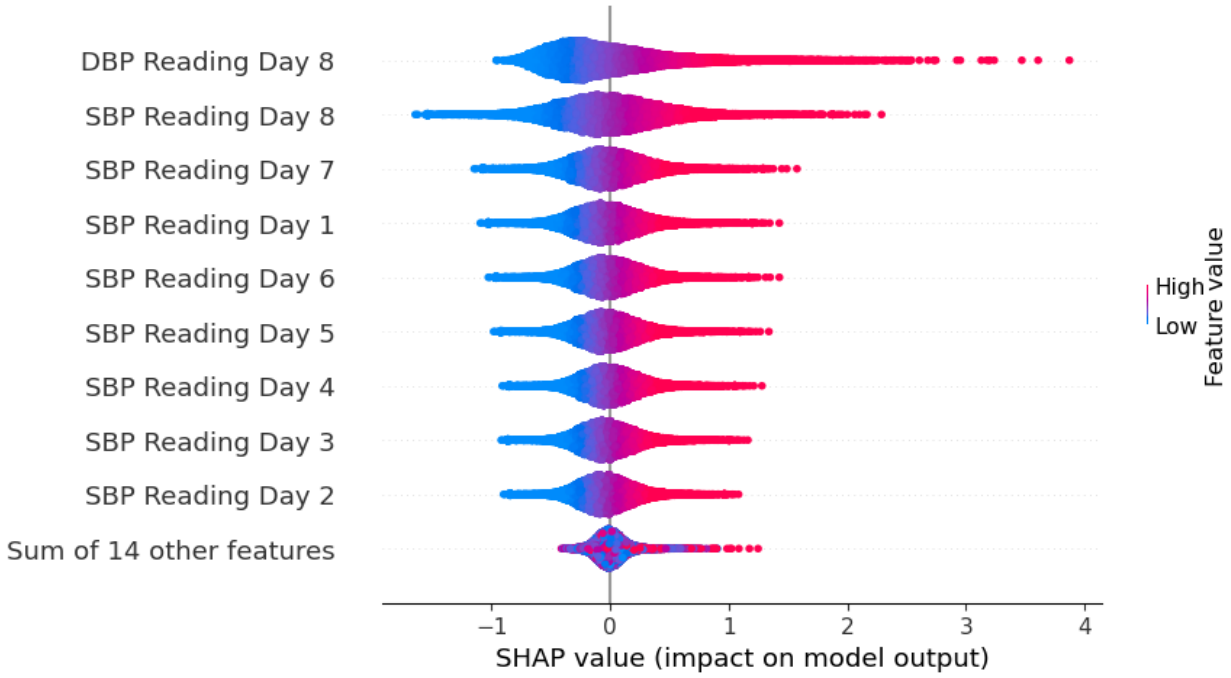

Figure 4.3: Logistic Regression SHAP plot with 7-day event horizon and 8-day sliding window.

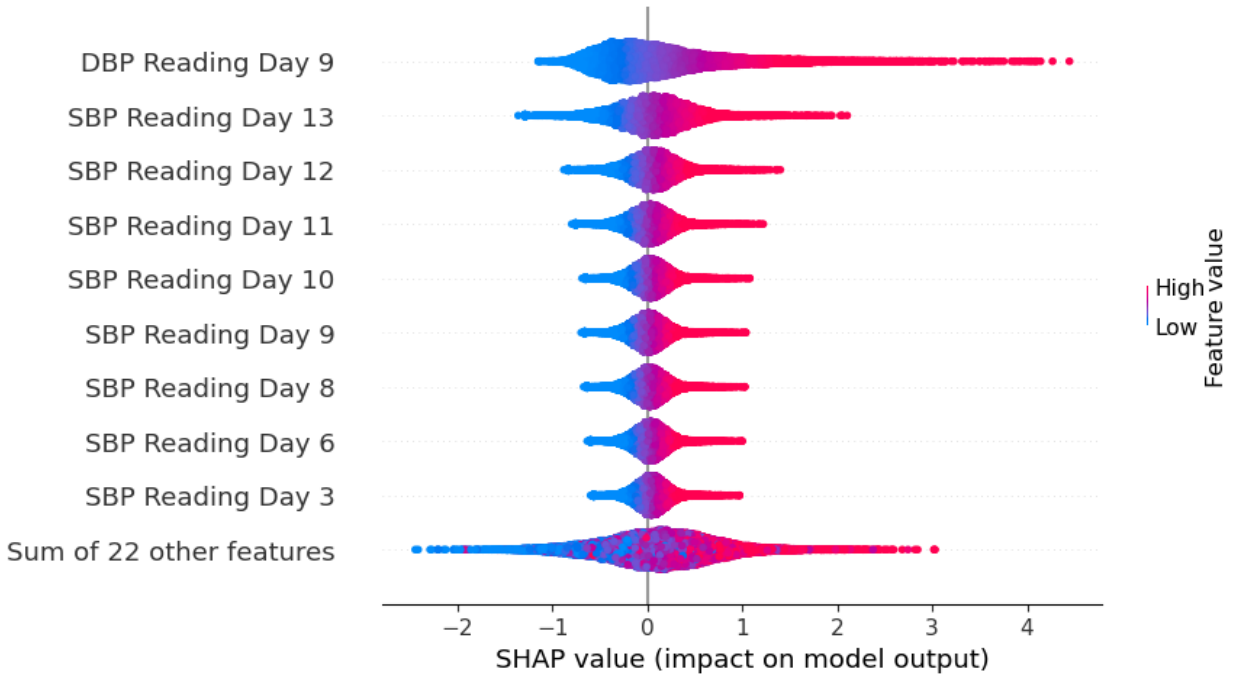

Figure 4.4: Logistic Regression SHAP plot with 7-day event horizon and 13-day sliding window.

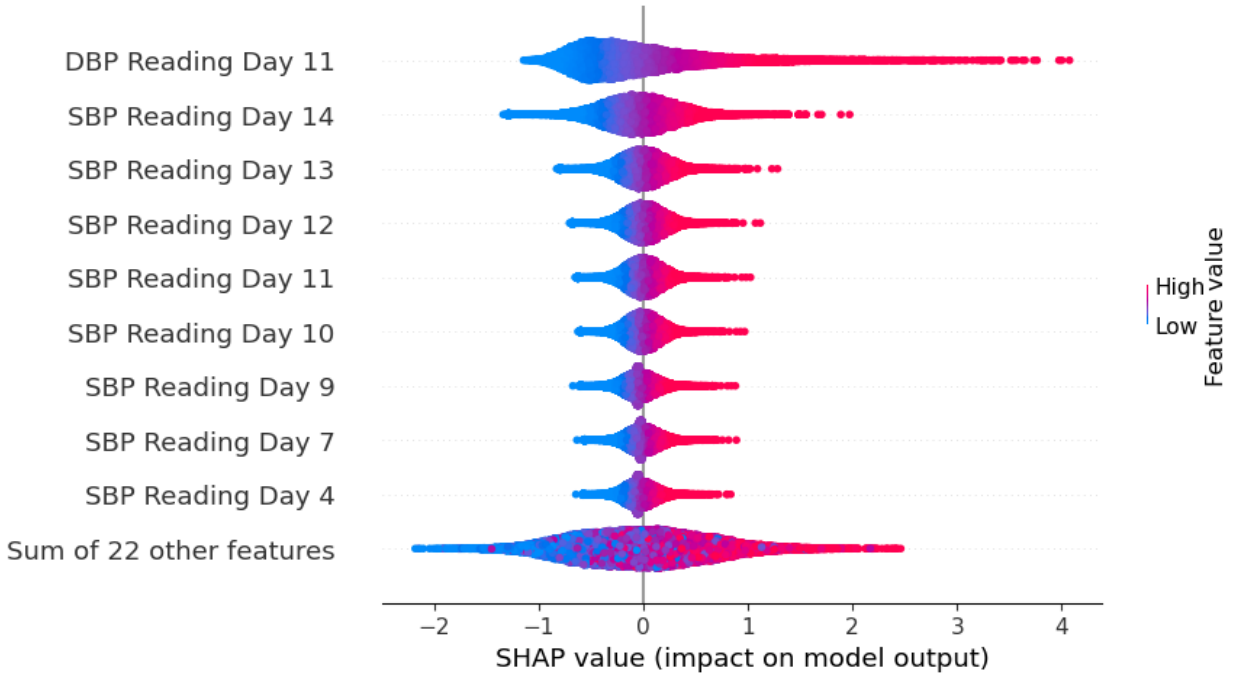

Figure 4.5: Logistic Regression SHAP plot with 7-day event horizon and 14-day sliding window.

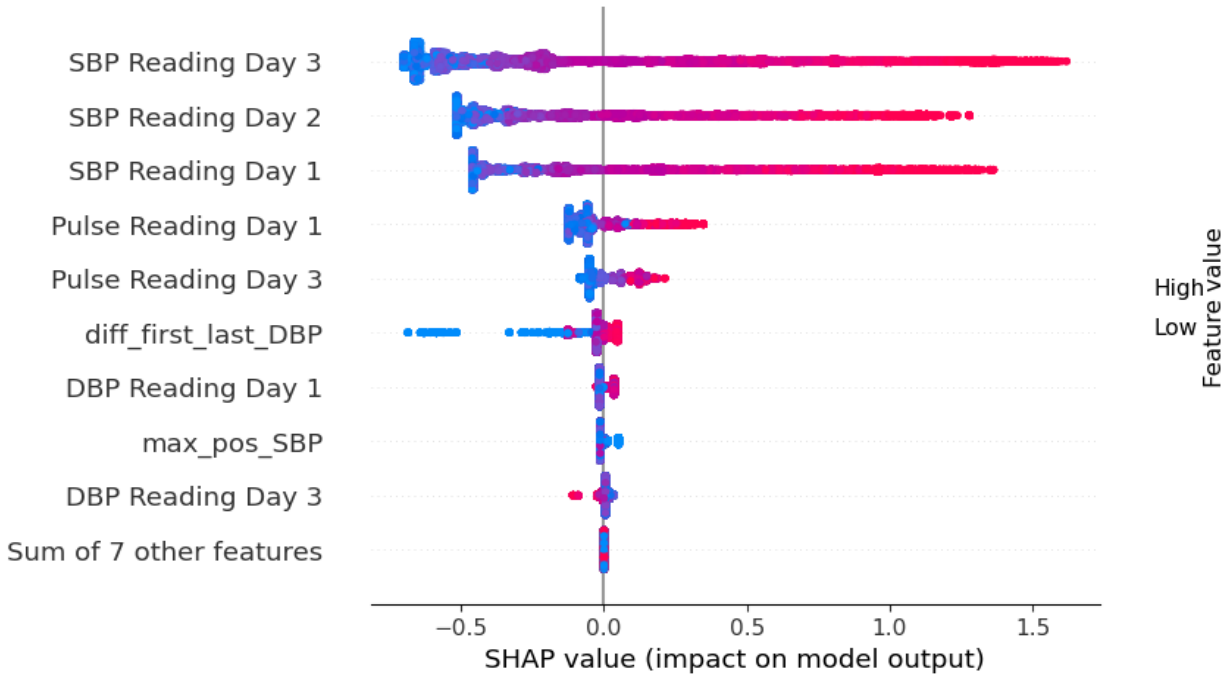

Figure 4.6: XG Boost SHAP plot with 7-day event horizon and 3-day sliding window.

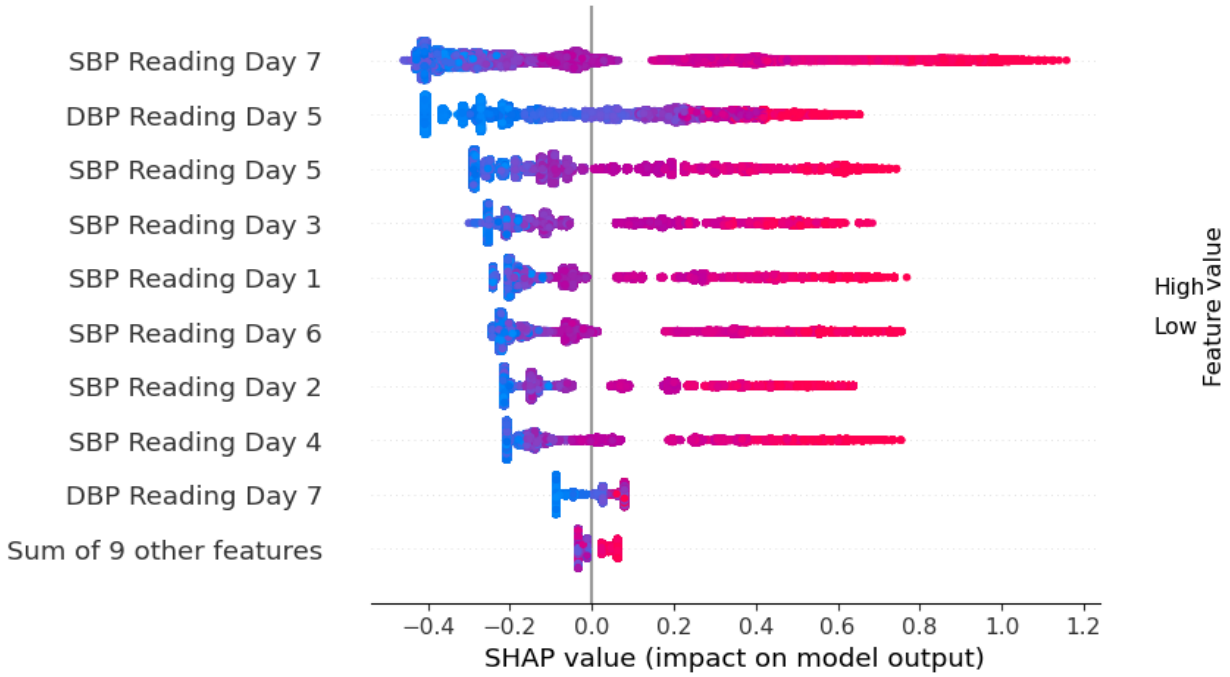

Figure 4.7: XG Boost SHAP plot with 7-day event horizon and 7-day sliding window.

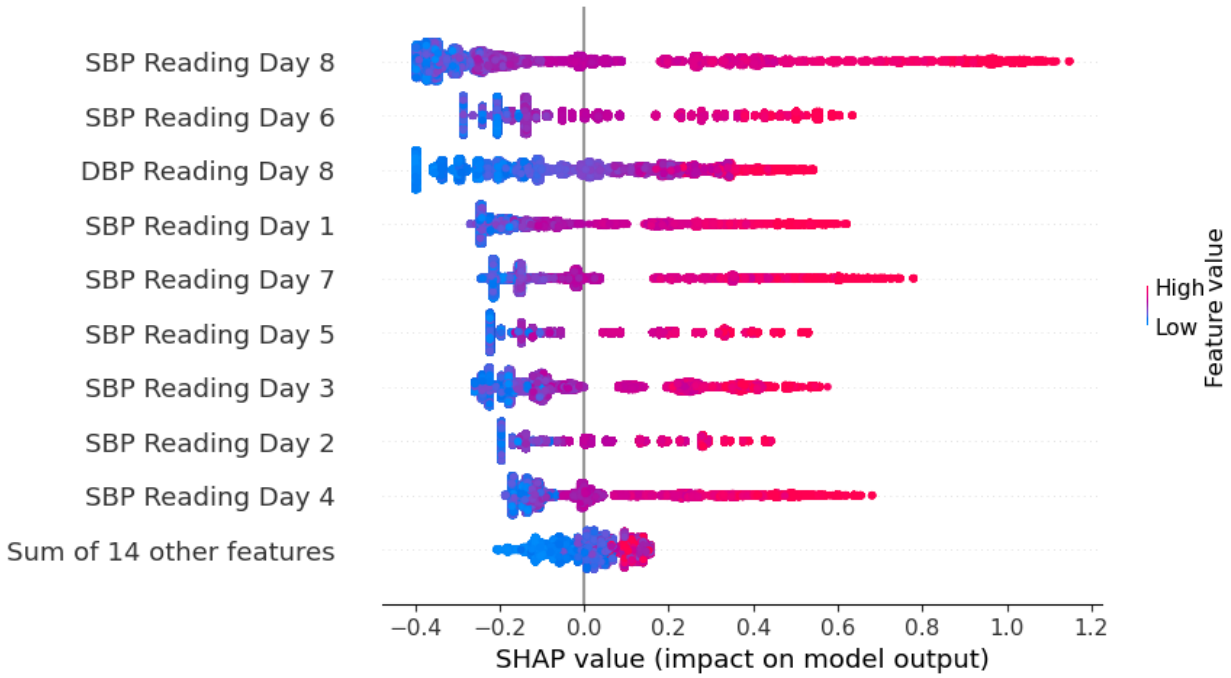

Figure 4.8: XG Boost SHAP plot with 7-day event horizon and 8-day sliding window.

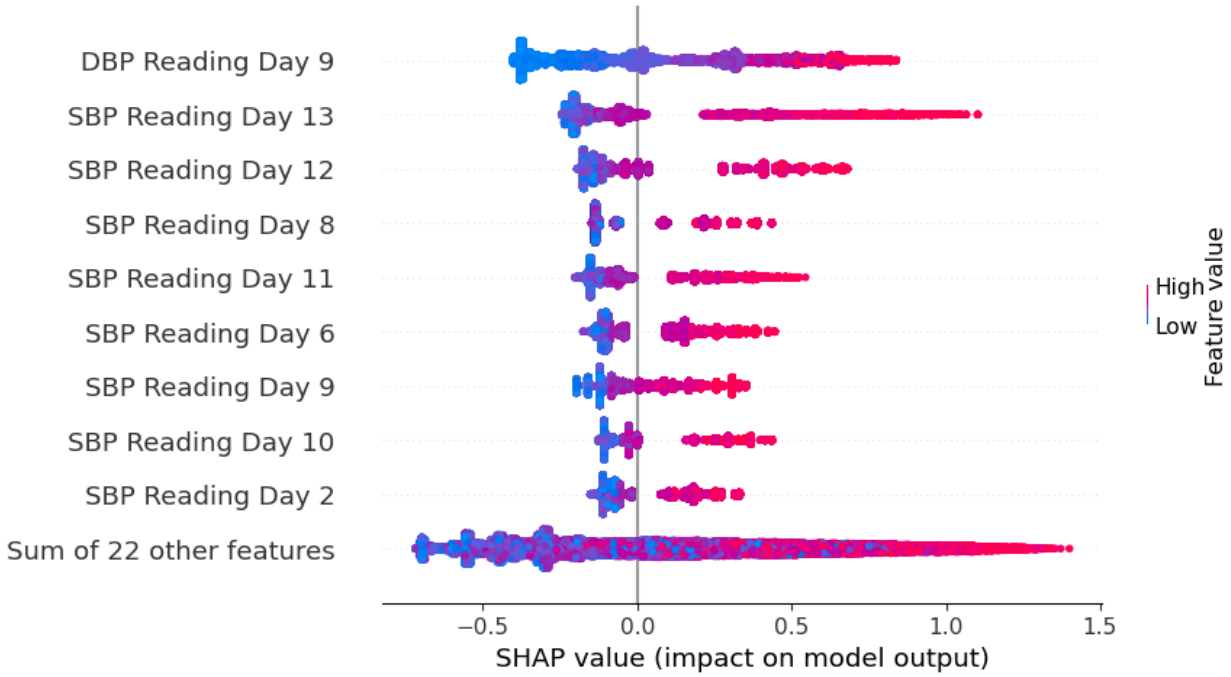

Figure 4.9: XG Boost SHAP plot with 7-day event horizon and 13-day sliding window.

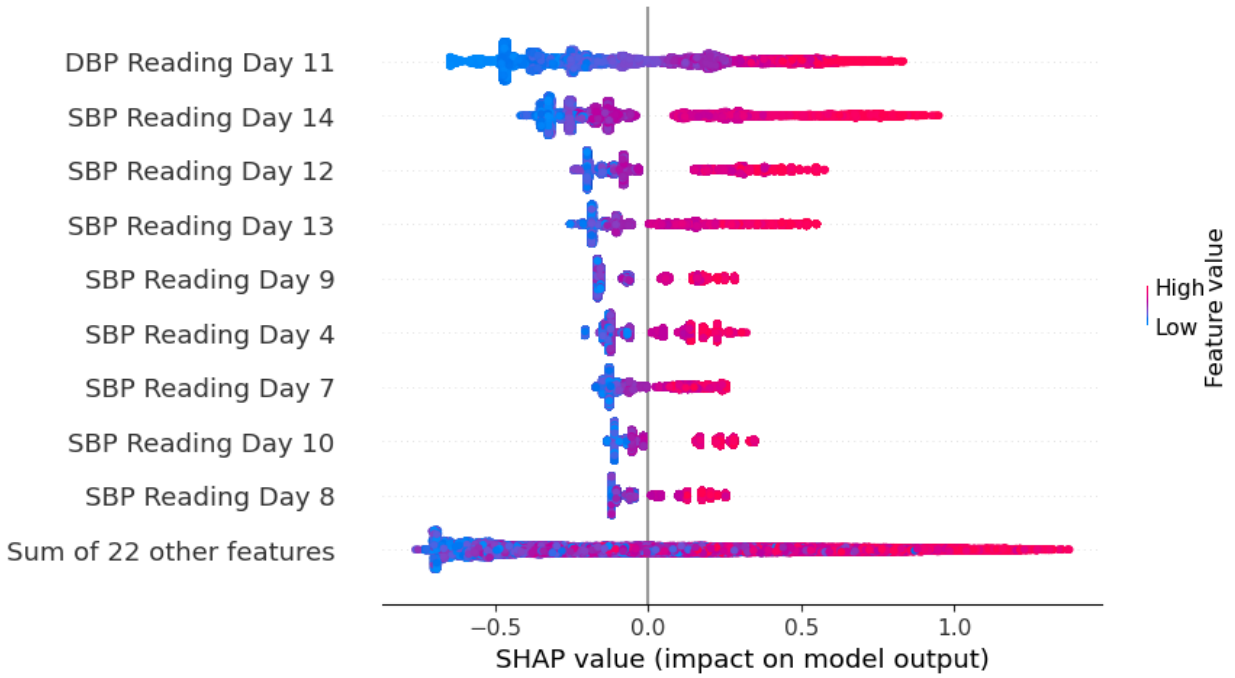

Figure 4.10: XG Boost SHAP plot with 7-day event horizon and 14-day sliding window.

4.1.3 Adaptive Model Overview

Using the best performing models for a 7-day event horizon as submodels to create the adaptive model results in the performance graphs shown by Figure 4.11 and 4.12. The model performances will change, depending on the event horizon and sliding window size chosen. As discussed earlier, the longer the event horizon, the better the classification performance. In the two Figures 4.11 and 4.12 we can also clearly see that while XG Boost does better than the other two classifiers when it comes to AUCROC values, this changes in the AUCPR figure. Here there is a more heterogeneous mix of classifiers with Logistic regression being better or equal to XGBoost on 10-day sliding windows, or 14-day sliding windows.

Figure 4.11: AUCPR values for each sliding window submodel in adaptive model

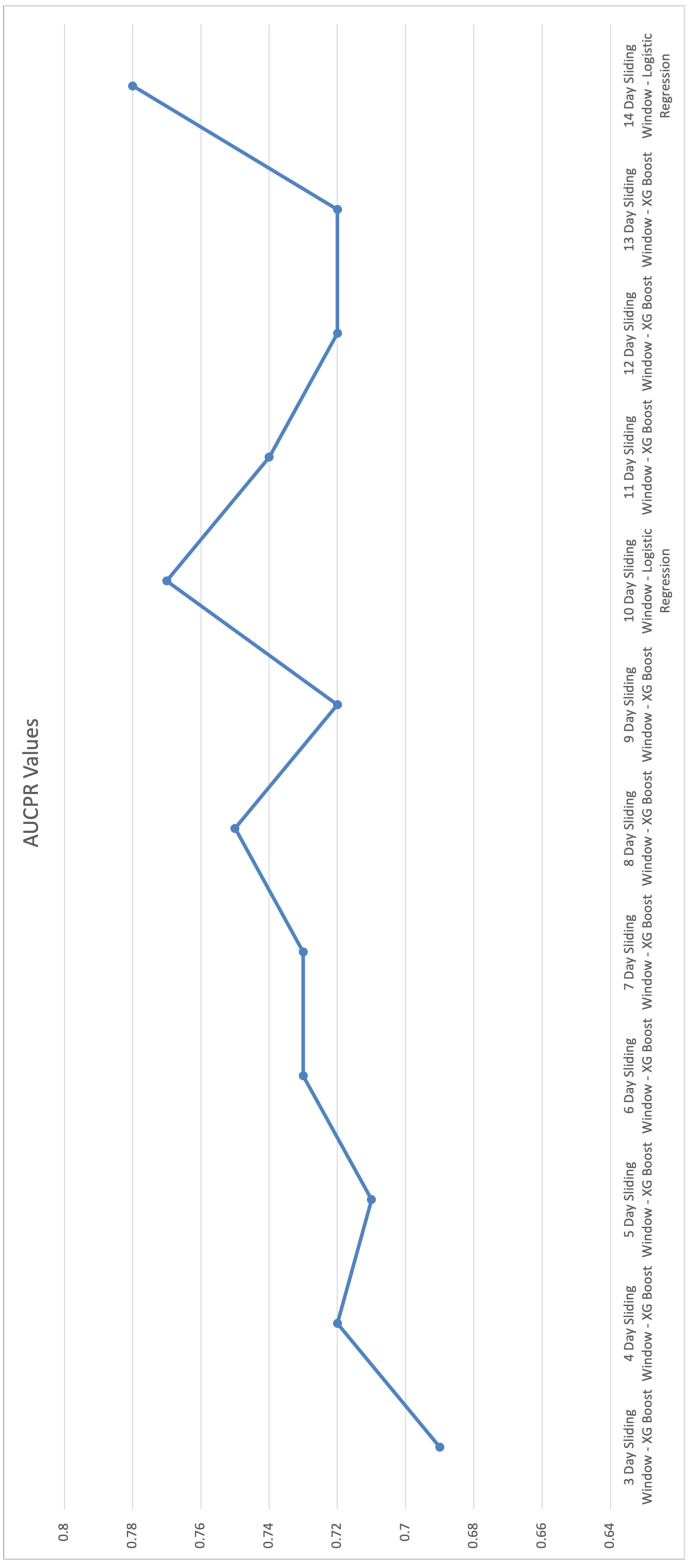

Figure 4.12: AUCROC values for each sliding window submodel in adaptive model

5. CONCLUSION & FUTURE WORK

5.1 Discussion

In general, all three models performed comparably, with Random Forest performance being slightly lower than Logistic Regression or XGBoost. This indicates that telemonitoring data does seem to generate data that has potential to generate alerts for preventive action. More importantly, the results indicate that the seven-day event horizon, since seven days, should provide sufficient time for potential intervention while retaining much of the accuracy. There are two factors that go into this result. First, in general, as the event horizon increases in size, performance improves because the algorithm does not have to be as precise on the exact day that a hypertensive event occurred (one-day horizon). However, the performances seem to start converging around a seven-day event horizon length. This may be because the features being used to predict are now further away from the event. That is, the last measurement used for prediction is seven days prior to the last day of the seven-day event horizon, but two weeks prior to the fourteen-day event horizon. Further, we can see that longer sliding windows yield higher accuracies. While the performance of the models increase as both the event horizon and the sliding window increase, they converge after a certain length. Looking at the Figure 4.2 we can see that the best performance was reached using a 13-day sliding window and 14-day event horizon before losing performance when increasing the sliding window size to 14 days. Since the performance of an 8-day sliding window is comparable to a 13-day sliding window model, in some cases even outperforming the latter, it might be more preferential to keep the sliding window sizes to around 7-8 days in length. More studies are needed to separate out these two orthogonal issues, namely how precisely can the algorithms reasonably predict and how many days out has optimal predictive performance.

The SHAP plot gives some additional insight into the second issue. In particular, the latest readings and the variance over the window matter, but the number of additional readings help the model better determine hypertension risk. The split in the feature ranges (SHAP value / along the

x-axis) in the XGBoost model demonstrate why the model is able to achieve a better AUCROC score, providing a likelihood that with increased data and increased participants, it will result in higher performance gains. Given the large overlap between the highest impacting features of both models, we demonstrate that our framework is able to successfully extract valuable features out of the available telemonitoring data and that additional days of data are important in accurately predicting periods of hypertension.

5.2 Limitations and Future Directions

This work presents promising ability to generate actionable alarms. However, the success of telehealth solutions is a three-step process: 1) finding the right timing for alarms to facilitate the right type of intervention, 2) keeping clinicians informed of any future event periods, and finally 3) evaluating those interventions for specific hypertensive events. In this paper, we address the first two questions with the limitation of only predicting periods of hypertension. A clear next step for this work is to evaluate the relationship between the alarms that would be generated from these models and a risk of hypertensive events, to then develop a future prospective trial. Additionally, the use of more advanced classification models such as deep neural networks or recurrent neural networks could enable prediction accuracy far exceeding those of the classifiers used in this study.

5.3 Conclusion

As previously established, hypertensive patients can be aided through remote health by daily monitoring of their SBP, DBP, and pulse. In this study, the patients' data is transmitted via Bluetooth to an intermediary device that connects to the cloud, enabling clinicians to view their patients' daily data. At the center of this software system is a framework which identifies key features, windows of measurement, and length of event horizon to successfully train a machine learning model to predict periods of hypertension. Additionally, an adaptive model enables physicians to begin using the framework early and relying on the best possible model for a given input sequence length. Based on the AUCROC and AUCPR of XGBoost ranging between 0.7-0.85 for AUCPR and 0.78-0.8 for AUCROC, we demonstrate the success of the framework in enabling prediction of periods

of hypertension. Throughout the process of aggregating, and then removing underperforming features, training a model, and then finally validating the model's performance using SHAP, we can see the highest impacting features are later in the sliding window. This indicates that, while the history of features matter, the latest trends are most indicative of positive events. This demonstrates the framework's ability to accurately extrapolate the most impactful features on the model output.

REFERENCES

[1] M. Birger, A. S. Kaldjian, G. A. Roth, A. E. Moran, J. L. Dieleman, and B. K. Bellows, "Spending on Cardiovascular Disease and Cardiovascular Risk Factors in the United States: 1996 to 2016," *Circulation*, vol. 144, pp. 271–282, July 2021.

[2] M. J. Stampfer and W. C. Willett, "Primary Prevention of Coronary Heart Disease in Women through Diet and Lifestyle," *The New England Journal of Medicine*, p. 7, 2000.

[3] S. E. Chiuve, T. T. Fung, K. M. Rexrode, D. Spiegelman, J. E. Manson, M. J. Stampfer, and C. M. Albert, "Adherence to a Low-Risk, Healthy Lifestyle and Risk of Sudden Cardiac Death Among Women," *JAMA*, vol. 306, July 2011.

[4] S. E. Chiuve, M. L. McCullough, F. M. Sacks, and E. B. Rimm, "Healthy Lifestyle Factors in the Primary Prevention of Coronary Heart Disease Among Men: Benefits Among Users and Nonusers of Lipid-Lowering and Antihypertensive Medications," *Circulation*, vol. 114, pp. 160–167, July 2006.

[5] S. E. Chiuve, K. M. Rexrode, D. Spiegelman, G. Logroscino, J. E. Manson, and E. B. Rimm, "Primary Prevention of Stroke by Healthy Lifestyle," *Circulation*, vol. 118, pp. 947–954, Aug. 2008.

[6] R. M. v. Dam, T. Li, D. Spiegelman, O. H. Franco, and F. B. Hu, "Combined impact of lifestyle factors on mortality: prospective cohort study in US women," *BMJ*, vol. 337, pp. a1440–a1440, Sept. 2008.

[7] S. Lewington, R. Clarke, N. Qizilbash, R. Peto, R. Collins, and Prospective Studies Collaboration, "Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies," *The Lancet*, vol. 360, pp. 1903–1913, Dec. 2002.

[8] P. Muntner, D. Shimbo, R. M. Carey, J. B. Charleston, T. Gaillard, S. Misra, M. G. Myers, G. Ogedegbe, J. E. Schwartz, R. R. Townsend, E. M. Urbina, A. J. Viera, W. B. White, J. T. Wright, and on behalf of the American Heart Association Council on Hypertension; Council on Cardiovascular Disease in the Young; Council on Cardiovascular and Stroke Nursing; Council on Cardiovascular Radiology and Intervention; Council on Clinical Cardiology; and Council on Quality of Care and Outcomes Research, "Measurement of Blood Pressure in Humans: A Scientific Statement From the American Heart Association," *Hypertension*, vol. 73, May 2019.

[9] M. Pasha, L. C. Brewer, S. Sennhauser, M. Alsawas, and M. H. Murad, "Health Care Delivery Interventions for Hypertension Management in Underserved Populations in the United States: A Systematic Review," *Hypertension*, vol. 78, pp. 955–965, Oct. 2021.

[10] S. S. Virani, A. Alonso, H. J. Aparicio, E. J. Benjamin, M. S. Bittencourt, C. W. Callaway, A. P. Carson, A. M. Chamberlain, S. Cheng, F. N. Delling, M. S. Elkind, K. R. Evenson, J. F. Ferguson, D. K. Gupta, S. S. Khan, B. M. Kissela, K. L. Knutson, C. D. Lee, T. T. Lewis, J. Liu, M. S. Loop, P. L. Lutsey, J. Ma, J. Mackey, S. S. Martin, D. B. Matchar, M. E. Mussolino, S. D. Navaneethan, A. M. Perak, G. A. Roth, Z. Samad, G. M. Satou, E. B. Schroeder, S. H. Shah, C. M. Shay, A. Stokes, L. B. VanWagner, N.-Y. Wang, C. W. Tsao, and On behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee, "Heart Disease and Stroke Statistics—2021 Update: A Report From the American Heart Association," *Circulation*, vol. 143, Feb. 2021.

[11] S. I. Chaudhry, J. A. Mattera, J. P. Curtis, J. A. Spertus, J. Herrin, Z. Lin, C. O. Phillips, B. V. Hodshon, L. S. Cooper, and H. M. Krumholz, "Telemonitoring in Patients with Heart Failure," *New England Journal of Medicine*, vol. 363, pp. 2301–2309, Dec. 2010.

[12] B. J. Mortazavi, N. S. Downing, E. M. Bucholz, K. Dharmarajan, A. Manhapra, S.-X. Li, S. N. Negahban, and H. M. Krumholz, "Analysis of Machine Learning Techniques for Heart Failure Readmissions," *Circulation: Cardiovascular Quality and Outcomes*, vol. 9, pp. 629–

640, Nov. 2016.

[13] M. K. Ong, P. S. Romano, S. Edgington, H. U. Aronow, A. D. Auerbach, J. T. Black, T. De Marco, J. J. Escarce, L. S. Evangelista, B. Hanna, T. G. Ganiats, B. H. Greenberg, S. Greenfield, S. H. Kaplan, A. Kimchi, H. Liu, D. Lombardo, C. M. Mangione, B. Sadeghi, B. Sadeghi, M. Sarrafzadeh, K. Tong, G. C. Fonarow, and for the Better Effectiveness After Transition–Heart Failure (BEAT-HF) Research Group, "Effectiveness of Remote Patient Monitoring After Discharge of Hospitalized Patients With Heart Failure: The Better Effectiveness After Transition–Heart Failure (BEAT-HF) Randomized Clinical Trial," *JAMA Internal Medicine*, vol. 176, p. 310, Mar. 2016.

[14] N. Alshurafa, C. Sideris, M. Pourhomayoun, H. Kalantarian, M. Sarrafzadeh, and J.-A. Eastwood, "Remote Health Monitoring Outcome Success Prediction Using Baseline and First Month Intervention Data," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, pp. 507–514, Mar. 2017.

[15] S. Park, H.-C. Kum, M. A. Morrisey, Q. Zheng, and M. A. Lawley, "Adherence to Telemonitoring Therapy for Medicaid Patients With Hypertension: Case Study," *Journal of Medical Internet Research*, vol. 23, p. e29018, Sept. 2021.

[16] S. Abrar, C. K. Loo, and N. Kubota, "A Multi-Agent Approach for Personalized Hypertension Risk Prediction," *IEEE Access*, vol. 9, pp. 75090–75106, 2021.

[17] J. Lee and R. Mark, "A Hypotensive Episode Predictor for Intensive Care based on Heart Rate and Blood Pressure Time Series," p. 4.

[18] D. Dave, D. J. DeSalvo, B. Haridas, S. McKay, A. Shenoy, C. J. Koh, M. Lawley, and M. Erraguntla, "Feature-Based Machine Learning Model for Real-Time Hypoglycemia Prediction," *Journal of Diabetes Science and Technology*, vol. 15, pp. 842–855, July 2021.

[19] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," p. 10.