# SOMATOSENSORY AMPLIFICATION AND FACIAL ACTION UNITS: A MACHINE LEARNING APPROACH TO PREDICT VASOVAGAL REACTIONS

MELIZ TYURKILERI

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

student number

2113955

committee

dr. Elisabeth M. J. Huis in 't Veld
Mr. Hosein Mohebbi

location

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

date

Jan 10th, 2025

word count

8777

acknowledgments

I would like to express my deepest gratitude to my thesis supervisor Dr. Elisabeth M. J. Huis in 't Veld for their invaluable insights and guidance throughout this journey. I am also sincerely grateful to the Faint Team (FAINT, 2018), for providing the data utilized in this study and making this project possible. I extend my heartfelt gratitude to my partner in crime, Esin Ulucaklı, as well as, my family and friends who have never second-guessed my abilities and dedication even when I have. Finally, I dedicate this work to all women who fight for their rights, freedom, and lives. In a world where access to education is still a privilege, I am aware of how lucky I am to come so far. However, I am committed to giving back and sharing the knowledge that I gained from this enhancing journey with others.

# SOMATOSENSORY AMPLIFICATION AND FACIAL ACTION UNITS: A MACHINE LEARNING APPROACH TO PREDICT VASOVAGAL REACTIONS

MELIZ TYURKILERI

**Abstract**

Blood donation procedures contain several stimuli such as needles, blood, or the experience of another donor that may trigger donors to have unpleasant feelings, namely vasovagal reactions. Experiencing vasovagal reactions may affect donors' retention behavior, posing a significant risk to sustainable blood supply. While self-reporting of vasovagal reactions may not allow early interventions, enhancing developments in video recordings and machine learning may enable early detection of these feelings. This study utilizes the intensities of facial action units of blood donors observed in earlier phases of donation, prior to needle insertion, and aims to classify donors based on their likelihood to experience low or high levels of vasovagal reaction during a donation. To this end, the study compares the three machine learning models trained: MLP, XGBoost, and Random Forest. The result reveals that Random Forest outperforms the remaining models with a more balanced and robust performance, attaining an F1 score of 0.49, a recall score of 0.55, and an accuracy score of 0.63. The study also assesses the performance of two different oversampling techniques: SMOTE and ROS, as well as, whether introducing somatosensory amplification to the feature set improves models' performances. The models with SMOTE demonstrate better performance compared to models with ROS, and joint utilization of somatosensory amplification and facial action unit intensities shows very limited and model-dependent improvement: only tree-based models exhibit a slight enhancement in prediction performance. The most informative facial expressions are identified as lip tightening, jaw drawing, and blinking. While the results of this study show room for improvement, it provides a promising baseline for future studies. The joint presence of the facial action units might be investigated in future studies, allowing more detailed mapping of underlying

emotions such as fear, pain, and anxiety. The scope of this study may also extended to a broader implication, ranging from mental health assessments to security.

# 1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The data utilized in this thesis has been acquired from the Faint Project (FAINT, 2018), in an anonymized format. The scope of this project did not include any data collection process from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. All the figures reported in this study were produced by the author. The replication package of the thesis has been shared through a publicly available GitHub repository, which can be accessed via the this link. The software and packages used to conduct the thesis are listed in Appendix A. ChatGPT (OpenAI, 2024) was utilized to improve the clarity of the author's original creation, and Grammarly (2024) was used for grammar and spell-checking. No other typesetting tools or services were used. A100 GPU provided by Google Colab (Google, 2024) was utilized for resource-demanding computational tasks and the diagrams provided in this study were designed by the author with the Lucidchart (Lucid Software, 2024).

# 2 INTRODUCTION

## 2.1 *Motivation and Social Relevance*

Access to a safe and sustainable blood supply is still challenging for many people across the world. According to the WHO, the number of blood donations reached 118.5 million in 2021 whereas 40 percent of worldwide donations have been collected from high-income countries. The regional gaps in blood donation have remained prominent, revealing that access to the supply of blood is still a risk factor specifically for the low and middle-income countries accounting for 84 percent of the global population (World Health Organization, 2022). Blood donations are the first step to ensuring a sustainable blood supply, which may contribute to saving numerous lives, especially in countries with limited resources.

While there are infrastructural and policy-relevant barriers such as lack of specific legislation, inadequate donation centers, insufficient inspection, and budgetary limitations (World Health Organization, 2022); individual factors, namely fear of needles (Ditto et al., 2010; France et al., 2013; Zucoloto et al., 2019), previous donation experience (France et al., 2005; Newman et al., 2006; Thomson et al., 1998), lack of information (Zucoloto et al., 2019) and donation intention (Huis In 't Veld et al., 2019) are also relevant to addressing challenges to a safe and sustainable blood supply.

Needle fear (Ditto et al., 2010; France et al., 2013) and previous donation experience (France et al., 2005; Newman et al., 2006; Thomson et al., 1998) were pointed out in earlier literature as challenges preventing individuals from becoming blood donors. These early empirical findings motivated the Faint Project (FAINT, 2018) aiming to develop an e-health game targeting individuals with needle phobia. Several studies (Rudokaite et al., 2024, 2023a; Rudokaite et al., 2023b) exploiting data from the Faint Project showed that machine learning algorithms can be effectively utilized to predict one of the adverse events, vasovagal reactions (VVR), that donors may experience during a donation process.

Motivated by the existing literature (FAINT, 2018), this study uses machine learning algorithms aiming to predict the likelihood of experiencing VVR during a blood donation. From a societal standpoint, these predictions could serve as an early warning mechanism for healthcare providers enabling timely interventions to enhance the donor experience, inherently contributing to a sustainable blood supply.

## 2.2 *Project Definition*

This thesis utilizes machine learning algorithms, namely Multi-layer Perceptron (MLP), XGBoost, and Random Forest, to predict the likelihood of experiencing VVR during a blood donation. It assesses the performances of the mentioned models which utilizes data on the donor's facial expressions up until a needle insertion phase. It exploits data on facial action units, a system developed by Ekman and Friesen (1976), which are primarily used to represent facial expressions. Somatosensory amplification as introduced by Barsky (1979) to assess which extent individuals heighten ordinary bodily sensations, is employed in this study as well.[1] To this end, the study compares the performance of the models trained independently on two different datasets: a baseline dataset only including facial action unit intensities and an extended dataset containing additional information on somatosensory amplification. Two alternative oversampling techniques; Synthetic Minority Over Sampling (SMOTE) and Random Over Sampling (ROS) also apply to each model on the baseline dataset, to underscore the comparative advantage of the different strategies to handle class imbalances. Then, the study concludes with an error analysis that underscores the limitations and improvement areas of the best-performing model and

[1] Facial action units and somatosensory amplification have been discussed extensively in further sections 3.1.1 and 3.1.2.

reports the results of a feature importance analysis which might serve as a foundation for future research.

## 2.3 *Related Work and Scientific Relevance*

### 2.3.1 *Machine Learning Algorithms for VVR Predictions*

Rudokaite et al. (2023b), extracted average temperature profiles from six facial regions prior to donation and then utilized these early temperature profiles to classify donors based on their likelihood to experience VVR during later phases of donation. They compared the performances of several algorithms, including Decision Tree, Random Forest, XGBoost, and Neural Network. XGBoost outperformed the other models with a higher recall and F1 score ( Recall=0.84, F1 score=0.86, PR-AUC=0.93). Their results revealed the importance of temperature changes in the nose, chin, and forehead area in predicting the risk of VVR.

Facial expressions were also used in the context of blood donation. Rudokaite et al. (2023a) employed XGBoost, Decision Tree, Random Forest, and Artificial Neural Networks (ANN) to predict vasovagal reactions based on intensities and the presence of facial action units from a pre-donation stage. The ANN outperformed the other algorithms with an F1 score of 0.82 (Recall=0.84, PR-AUC=0.79), standing out as the state-of-the-art to predict VVR based on the intensities of facial action units. Analysis of features highlighted the importance of the intensities in the eye region.

Another promising study (Rudokaite et al., 2024) was conducted to exploit continuous video recordings from the pre-donation stages to classify donors based on VVR risk groups. This study also compared the performance of machine learning models across different datasets and found that the pre-trained ResNet152 models with GRU on continuous video data outperformed the other models (Recall=0.58, F1=0.69, PR-AUC=0.81). The findings underscored the importance of eyes, lips, chin, and forehead in these predictions.

### 2.3.2 *Scientific Relevance*

While there are promising efforts to improve VVR predictions by utilizing machine learning algorithms on facial data, they (Rudokaite et al., 2024, 2023a; Rudokaite et al., 2023b) exclusively used data on prior-to-donation, covering a period before the act of blood donation actually began. These studies utilized a subset of the data collected within the scope of the Faint Project (FAINT, 2018), still a significant portion of the available data remained uncovered.

From a scientific perspective, this thesis contributes to the relevant literature by extending the scope of facial data utilized in VVR predictions. Given the fact that different stages of donation have distinct psychological and physical stimuli, incorporating data from the earlier stages of donation, a phase more pronouncedly including physical stimuli (such as the sight of a needle, tourniquet, or sitting at donation chair), may enhance the information derived from facial expressions. [2] To this end, it differentiates itself from previous literature primarily by exploiting data on donors' facial expressions up until the needle insertion phase, including an earlier stage after the donation process actually begins.

## 2.4 *Research Strategy*

VVR associated with blood donation relied on many risk factors such as gender, age, anxiety, and expected pain (Thijsen & Masser, 2019). However, observing all factors is not possible due to resource constraints, and measuring these factors is also challenging. Besides, unrevealed patterns between them may jointly exacerbate the risk of experiencing VVR.

Multi-layer Perceptron (MLP), a type of ANN, provides a powerful architecture to study non-linear complex relationships (Goodfellow et al., 2016). An MLP architecture might cover complex interactions between the facial micro-expressions. However, utilizing an MLP could be challenging given the limited data size. Therefore, comparing MLP with other machine learning algorithms is essential to quantify its performance. This study assesses the performance of MLP with the other two algorithms XGBoost and Random Forest. XGBoost and Random Forest utilized by the previous studies (Rudokaite et al., 2024, 2023a; Rudokaite et al., 2023b) are selected in this research considering their ability to capture non-linear relationships and to perform well even with small datasets.

In line with the discussion above and the identified knowledge gap in current literature (see section 3.1.3), the main research question and the sub-questions of this thesis are outlined as follows:

- **Main Research Question:** Does Multi-layer Perceptron (MLP) perform better than XGBoost and Random Forest in predicting vasovagal reactions, based on donors' facial action unit intensities extracted until a needle insertion?

- **SRQ1:** Does MLP outperform the XGBoost, and Random Forest by reaching a higher F1 score?

[2] A more detailed discussion on this matter is provided in section 3.1.3.

The dataset of this study suffers from a class imbalance. While the Synthetic Minority Oversampling Technique (SMOTE) performs well with the dataset in hand (Rudokaite et al., 2023a), it is computationally expensive compared to Random Oversampling (ROS). This study compares the two alternative oversampling methods to quantify the added value of SMOTE.

- **SRQ2:** Can utilizing ROS improve models' PR-AUC and F1 scores compared to employing SMOTE?

Somatosensory amplification accounting for individual variations in perception of internal and external threats to body integrity (Köteles & Witthöft, 2017), may help to distinguish between low and high VVR groups. To this end, this study jointly utilizes the total SSAS score along with facial action units.

- **SRQ3:** Does adding a somatosensory amplification scale to the feature set further improve the F1 score of the best model?

- **SRQ4:** Does the prediction performance of the best model differ across the donor groups: **[a]** donors who experienced VVR in previous blood donation, **[b]** donors who have never experienced VVR during blood donation, **[c]** new donors?

# 3 LITERATURE REVIEW

## 3.1 *Vasovagal Reactions (VVR)*

Improving the blood donation experience is an effective way to ensure a sustainable blood supply. Studies on blood retention behavior reveal that donors who experience adverse reactions associated with the donation are less likely to be donors again (France et al., 2013; Newman et al., 2006; Thijsen et al., 2019). Vasovagal reactions (VVR), with a changing severity from sweating to fainting, are the most common adverse event that blood donors experience (Garozzo et al., 2010; Kumari, 2015).

Several factors jointly increase the risk of experiencing VVR: observable donor characteristics (gender, age, donation history), unobservable characteristics (less sleep duration, anticipated pain, greater anxiety, history of VVR), and contextual characteristics (longer wait and bleeding time, witnessing VVR) (Thijsen & Masser, 2019). Indeed unobservable characteristics such as; fear of blood (Ditto et al., 2012; France et al., 2019; Gilchrist et al., 2015), needle fear (France et al., 2019; Hamilton, 1995; Sokolowski et al., 2010), pre-donation anxiety, expected pain and anxiety (Olatunji et al., 2010), are commonly pointed out in the current literature as the drivers of VVR.

Previous literature (Almutairi et al., 2017; Wang et al., 2019) extensively relied on observable characteristics such as gender, age, and weight, as well as linear models, to assess predictors of VVR. While these studies serve as a point of origin, they did not account for unobservable factors; which may be better proxies to predict VVR (Meade et al., 1996). Fortunately, recent improvements in the image and video processing techniques have enabled studies (Rudokaite et al., 2024, 2023a; Rudokaite et al., 2023b) to incorporate unobservable risk factors in VVR predictions.

### 3.1.1 *Facial Action Units*

The Facial Action Coding System (FACS) proposed by Ekman and Friesen (1976) introduced a new dimension to the literature by allowing it to detect and distinguish all visible facial behaviors. Recent studies relied on the FACS and facial expressions proved that the machine learning algorithms are safely utilized to predict stress (Blasberg et al., 2023; Giannakakis et al., 2022; Viegas et al., 2018), anxiety (Gavrilescu & Vizireanu, 2019) and pain (Bargshady et al., 2020; Hammal & Cohn, 2012; Werner et al., 2014).

### 3.1.2 *Somatosensory Amplification*

Somatosensory amplification introduced by Barsky (1979); is characterized as a tendency to heighten normal bodily sensations and experience them as more intense and disturbing (Barsky & Silbersweig, 2023). Several studies adopted the Somatosensory Amplification Scale (SSAS) (Barsky et al., 1990) to measure its association with anxiety (Barsky et al., 1988), pain (Köteles & Witthöft, 2017), and physical functioning (Barends et al., 2020). The SSAS was also used to predict VVR (Rudokaite et al., 2024). However, its joint effect along with the facial action units has not been studied before.

### 3.1.3 *State of the Art and Knowledge Gap*

Rudokaite et al. (2023a), previously mentioned in section 2.3.1 establishes a robust benchmark in VVR predictions exploiting facial expressions. They trained a set of algorithms - ANN, Random Forest, XGBoost, Decision Tree- on three distinct datasets, each focusing exclusively on one aspect: pre-donation VVR ratings, the intensity of facial action units, and the presence of facial action units. ANN model on the intensity of facial action units was revealed as the best-performing model (Recall=0.84, Precision=0.79, F1=0.82, PR-AUC=0.79), while Random Forest on the same dataset ranked as the second-best performing alternative (Recall=0.84, Precision=0.79, F1=0.82, PR-AUC=0.79) among the 12 model-dataset pairs utilized. The ANN model on the intensity of facial action units stands out as the state-of-the-art in VVR predictions leveraging facial expressions. These results are particularly important for the literature since they demonstrate that the dependence of VVR predictions on self-reported data might decrease in the presence of data on facial expressions.

While this thesis shares similarities with the state-of-art (Rudokaite et al., 2023a) in certain ways, it also exhibits distinct characteristics that set it apart. Although, Rudokaite et al. (2023a) used machine learning models to predict vasovagal reactions based on the donors' facial micro-expressions they exploited the data only from the pre-donation process while the donors were still waiting for the donation in the waiting room. This study provides promising and motivating results, yet it does not utilize all the available data. They extracted information using 2 to 3-minute video recordings corresponding to the first two stages, remaining a significant part of the available data uncovered. The novel data collected by the Faint Project (FAINT, 2018) extends beyond the pre-donation stages and tracks donors throughout the donation procedure. Within the scope of the Project, continuous video recordings spanning 5 to 20 minutes were collected from donors[3], including the whole period when they are sitting in a donation chair and waiting for donor assistants to prepare the needle.

Incorporating data from the earlier stages of donation (after donation actually starts) might be a noteworthy effort to improve VVR predictions, as each stage has distinct stimuli. While the pre-donation stages mostly include psychological stimuli such as anticipated pain, fear, and anxiety; the later stages primarily involve physical stimuli such as the sight of needles and tourniquet, sitting at a donation chair, watching a donor assistant during preparation, or smells of medical equipment. Different stimuli may trigger varied responses associated with VVR and alter the information conveyed by facial action units across stages.

Considering the discussion above, this study utilizes data on the facial micro-expressions not only before donation but also up to the point where a needle is inserted, which has not been studied before. This thesis aims to narrow existent knowledge gaps in the current literature by exploiting data from not only pre-donation but also an early stage of the donation process (up until a needle insertion) including both psychological and physical stimuli, as being a first study that jointly utilizes somatosensory amplification and intensities of facial action units, and by incorporating more observations which is particularly important considering the class-imbalance in-hand.[4]

[3] These continuous recordings cover stage 4 to stage 6

[4] The coverage of the Faint Dataset increased throughout the period. Although Rudokaite et al. (2023a) used data on 227 actual blood donors, this study exploits information on 306 actual blood donors.

# 4 METHODOLOGY

## 4.1 *Data*

### 4.1.1 *Source*

The Faint Project (FAINT, 2018) collected data from 328 blood donors at Sanquin, the national blood bank of the Netherlands. Within the scope of the Project, the donation procedure was designed in 8 stages described in Figure 1. At the baseline, the donors filled out a personality questionnaire, reporting information on demographics (gender, age), needle fear, pre-donation history, and the somatosensory amplification scale. Moreover, the donors were recorded using a digital camera (Nikon Coolpix AW130) throughout the donation procedure from stage 1 (after questionnaire) to stage 7 (at the donor cafe). Self-reported VVR scores were also collected from the donors in each donation stage. The Faint Dataset (FAINT, 2018; Rudokaite et al., 2023a), is a novel data source to study VVR since it allows extraction of facial action unit intensities that may proxy unobserved risk factors of VVR.

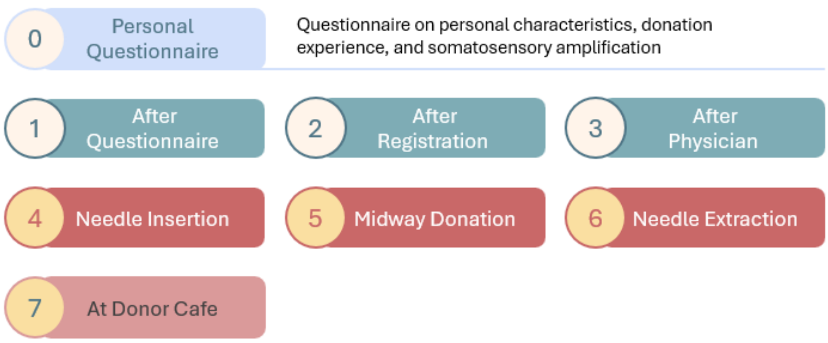



Figure 1: Stages of the donation procedure

This study utilizes a preprocessed version of the Faint Dataset. It incorporates data on facial action unit intensities for each donor, questionnaire responses containing information on donor characteristics, self-measured VVR scores, somatosensory amplification scale, and timestamps showing the starting and ending points of donation stages for each donor.

### 4.1.2 *Origin of Features*

This section provides an overview of the variables reported in the Faint Dataset and forms the basis of the features used in this study.

**Intensities of facial action units:** Video recordings collected within the scope of the Faint Project were pre-processed by the Faint Team. They utilized the OpenFace (Baltrusaitis et al., 2018) to extract the intensities of 17 distinct action units listed in Table 1. In the final version of the OpenFace outputs, the intensity levels are represented on a continuous scale, ranging from 0 to 5. Higher values indicate the more intense activity level for a given action unit (AU).

**Self-reported VVR scores:** In the data collection process conducted by the Faint team, donors were asked in each stage to what extent they experienced psychological (tension, fear, stress, and nervousness) and physiological ( weakness, dizziness, faintness, and light-headedness) reactions. A 5-point Likert scale (1= not at all, and 5= extremely) was used to collect self-reported VVR scores from the donors.

Table 1: Facial action units

| AU Code | Name |
|---|---|
| AU1 | Inner Brow Raiser |
| AU2 | Outer Brow Raiser |
| AU4 | Brow Lowerer |
| AU5 | Upper Lid Raiser |
| AU6 | Cheek Raiser |
| AU7 | Lid Tightener |
| AU9 | Nose Wrinkler |
| AU10 | Upper Lip Raiser |
| AU12 | Lip Corner Puller |
| AU14 | Dimpler |
| AU15 | Lip Corner Depressor |
| AU17 | Chin Raiser |
| AU20 | Lip Stretcher |
| AU23 | Lip Tightner |
| AU25 | Lips Part |
| AU26 | Jaw Drop |
| AU45 | Blink |

*Note.* Table 1 is the modified version of Table 1, originally presented by Ekman and Friesen (1976).

**Donor type:** Donors were grouped into three subcategories in the original dataset based on their previous donation histories: **[a]** donors who experienced VVR in previous blood donation, **[b]** donors who have never experienced VVR during blood donation, and **[c]** new (first-time) donors.

**Somatosensory amplification scale (SSAS):** Before the donation process started, the donors were asked to what extent they heighten normal bodily sensations and experience them intensely. A 10-item questionnaire shown in Table 2 developed by Barsky et al. (1990) was used to measure somatosensory amplification. Donors responded on a 5-point Likert scale for each item, with values ranging from 1 (not at all) to 5 (extremely).

Table 2: Somatosensory amplification scale

| Item |
|---|
| 1. When someone else coughs, it makes me cough too |
| 2. I can't stand smoke, smog or pollutants in the air |
| 3. I am often aware of various things happening within my body |
| 4. When I bruise myself, it stays noticeable for a long time |
| 5. Sudden loud noises really bother me |
| 6. I can sometimes hear my pulse or my heartbeat throbbing in my ear |
| 7. I hate to be too hot or too cold |
| 8. I am quick to sense the hunger contractions in my stomach |
| 9. Even something minor, like an insect or a splinter, really bothers me |
| 10. I have a low tolerance for pain |

Note: Table 2 is the modified version of Table 1, originally presented by Barsky et al. (1990).

### 4.1.3 *Preprocessing*

Facial action unit intensities extracted by the Faint team are provided as distinct data files corresponding to each donor's different stages of the donation procedure. As a first step of the preprocessing, the data files were merged to create a master dataset reporting facial action unit intensities for each donor across all the relevant stages. Data files including stage 4 (needle insertion) were filtered before the merging process as seen in Figure 2. The timestamp data, reporting the starting and ending time points of stages for each donor, was used to extract relevant information from the continuous data file covering stages 4 to 6. [5]

The timestamps of the master dataset were re-coded to ensure a continuous data flow of facial action unit intensities belonging to each donor[6] The final action unit dataset contained information on 3,362,818 frames be-

[5] The master AU dataset was restricted to the timeframes up to the needle insertion point, excluding stage 3. The stages of the donation procedure were shown slight variations across the blood collection centers, and stage 3 (after physician) was skipped for a subset of donors. Since the video recordings from this stage are only available for some donors, it is excluded.

[6] This step was necessary because the timestamps in the original data files reset to zero at the start of each stage.

longing to 313 blood donors. Then, the Tsfresh (Christ et al., 2018) python package was utilized to extract time series characteristics of the facial action units. For each of the 17 action units listed in Table 1, 9 summary characteristics (including total sum, variance, standard deviation, mean change, mean, median, minimum, maximum, and root mean square) were extracted. Lastly, the extracted time series characteristics were merged with the questionnaire data containing information on demographics, previous donation experiences, and responses to the somatosensory amplification scale questions. This merged dataset reporting information on 328 blood donors, was utilized in the data cleaning and exploratory data analysis steps.

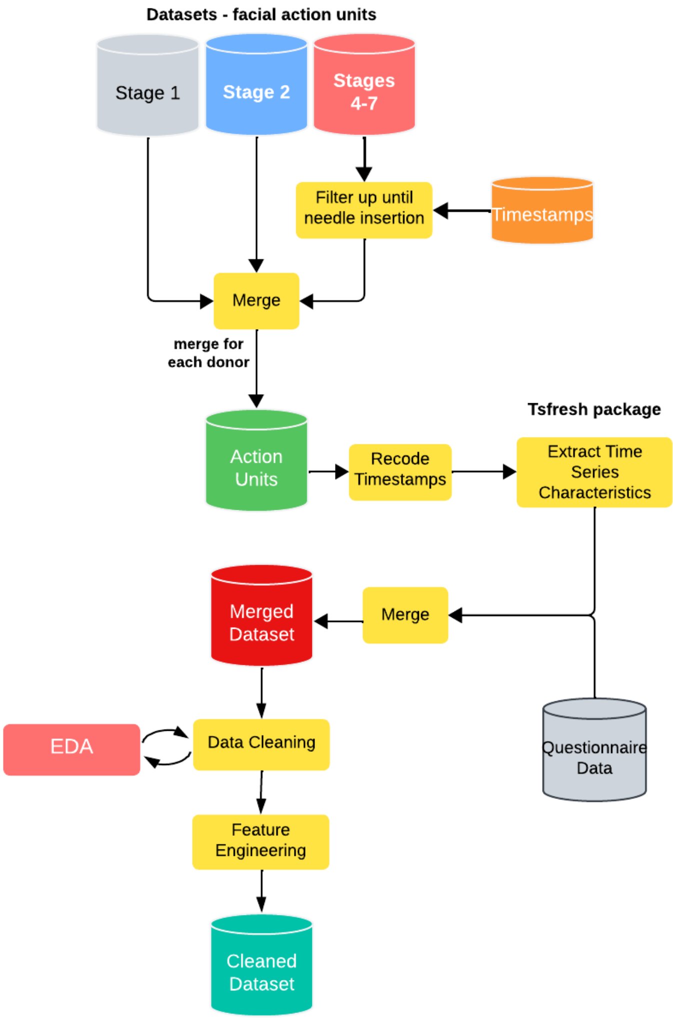



Figure 2: Overview of preprocessing

### 4.1.4 *Data Cleaning*

After pre-processing, the merged dataset was investigated to understand the missing values and three broader groups were identified: a. observations that have missing values in at least one VVR score, but no missing values in rest (5 observations), b. observations that have missing values in at least one AU unit or sub-items of SSAS, but no missing values in VVR scores (16 observations), c. observations that have missing values both in VVR scores and other characteristics (SSAS, donor type, etc.), but no missing values in action units (1 observation).

Category a contains donors whose procedure was deferred due to travel or low hemoglobin. Since the facial action units were extracted up until a needle insertion phase, corresponding values were not missing for this group. However, they could not complete the donation. Likewise, the observation reported under group c could not finish the procedure. These observations were excluded, and the sample was restricted to donors who completed the donation.

While the decision on categories a and c was relatively straightforward, it was not the case for donors pertained to category b. These observations demanded a more meticulous examination. The missingness pattern of them shows a systematic difference across donor types. These patterns reveal that being first-time donors increases the likelihood of having missing values in the time series characteristics of action units. This finding ruled out the missing completely at random (MCAR) mechanism. Then, the Mann-Whitney U test was applied to assess the difference in total VVR score distributions based on a binary indicator for missing cases (category b). The result was insignificant (p-value=0.397), suggesting that the missingness is less likely to demonstrate a missing not-at-random (MNAR) pattern.

A list-wise deletion might introduce a bias in missing at random (MAR) and MNAR scenarios,still, it needs to be considered in this case. Time series characteristics were not extracted for these observations since the video recordings collected for the Faint Project were not available for the mentioned donors. To this end, implementing more advanced techniques such as multiple imputation or Bayesian modeling might still be problematic. Although there are non-missing characteristics such as gender, age, and type of donor which might be used only for the imputation step, 153 variables(time characteristics of action units), a whole feature set, still need to be imputed under this scenario, which might result in noisier estimates. Since the share of observations under category b represents only 5 percent of the total sample, the decision has been made in favor of likewise deletion. The cleaned sample was restricted to 306 observations.

### 4.1.5 *Feature Engineering*

The feature engineering steps implemented within the scope of the project are listed below.

**SSAS score:** SSAS score was defined as the sum of responses to the 10-item somatosensory amplification scale reported in Table 1.

**MinMaxScaler:** MinMaxScaler allowing the transformation of numeric features was applied for normalization. The numeric features were mapped to the 0-1 ranges which is specifically important for the SVM and MLP algorithms that are sensitive to feature scales. As shown in Figure 7, the normalization step was applied as a part of the pipeline to prevent data leakage, and the normalization was skipped when Random Forest or XGBoost was utilized in a given pipeline.

### 4.1.6 *Labels*

As described in section 4.1.2, The Faint Team has collected self-reported psychological and physiological VVR scores from donors at each stage. The corresponding scores from stages 4 to 7 have been utilized in this study as a proxy for experiencing VVR in the later stages of donation. To validate the internal consistency of these sub-scores as a measure of the same underlying concept, a Cronbach's Alpha test was applied. Given the acceptable internal consistency ( alpha=0.782, 95% confidence interval: 0.74-0.82)[7], the total VVR score was defined as the sum of the 8 different sub-scores. This approach simplified the classification task at hand by enabling a binary classification, rather than designing the study as a multi-class classification problem. A multiclass classification problem might be particularly challenging and redundant in this case, considering the small sample size, class imbalances, and the fact that the sub-scores measure the same underlying concept.

To this end, the mean of the total VVR score was used as a threshold to classify donors based on VVR groups[8]. The donors with total VVR scores higher than the threshold were clustered as the high VVR group, while the remaining donors were addressed as the low VVR group. This group definition was utilized as the label of the binary classification problem. The final sample consists of 207 donors from the low VVR group, and 99 donors from the high VVR group. Figure 3 demonstrates the class imbalance in the final sample, emphasizing the necessity of employing an oversampling technique.

---

[7] An alpha value above 0.7 is generally considered as a threshold to validate internal consistency.

[8] A further discussion on the utilization of mean scores is provided in section 6.2

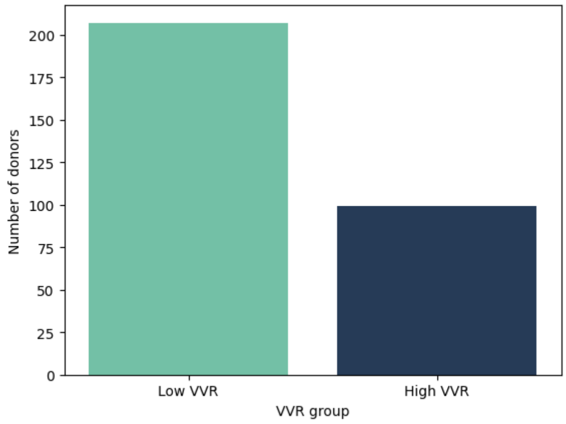



Figure 3: Distribution of classes

### 4.1.7 *Exploratory Data Analysis*

Exploratory data analysis (EDA) was not separated from the previous steps described above. Mainly, it was positioned as an activity that enhances the decisions made during the data cleaning and feature engineering processes.

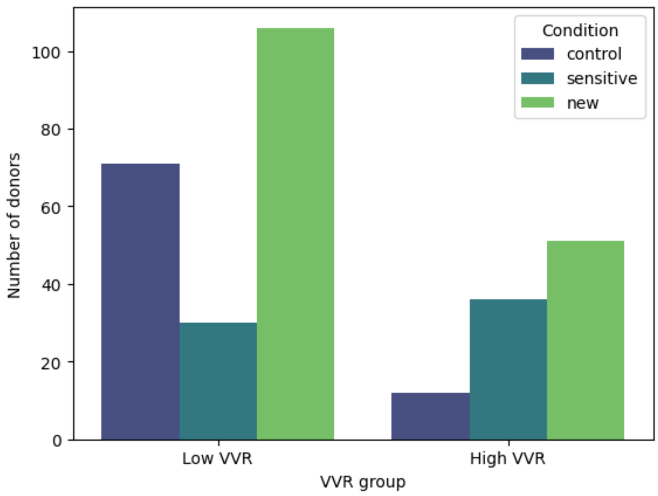



Figure 4: Distribution of classes by donor type

The final sample includes 306 donors, comprising 179 females and 127 males. The observations are represented under three distinct groups according to the previous donation histories as described in section 4.1.1: Control group (n=83), Sensitive group (n=66), and New donors (n=157). Figure 4 illustrates the composition of high and low VVR groups by donor type. The majority of the control donors are classified under the low VVR group, whereas the sensitive donors have a higher representation in the high VVR group. The final sample includes a higher proportion of females, where the gender distribution is more uneven in the higher end of the VVR distribution (see Figure 5).[9]

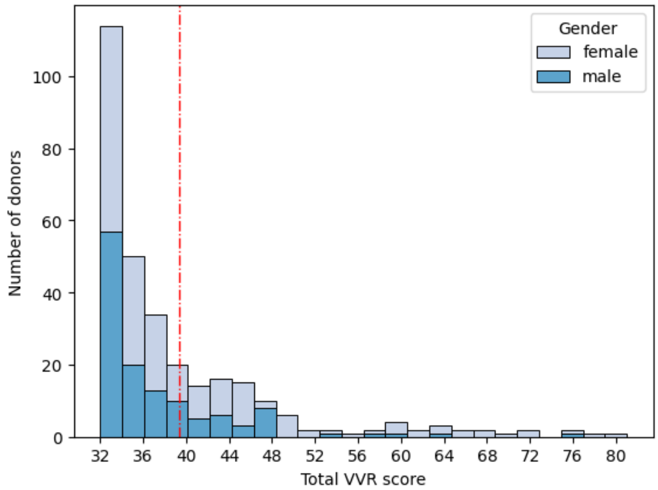



Figure 5: Distribution of total VVR scores by gender

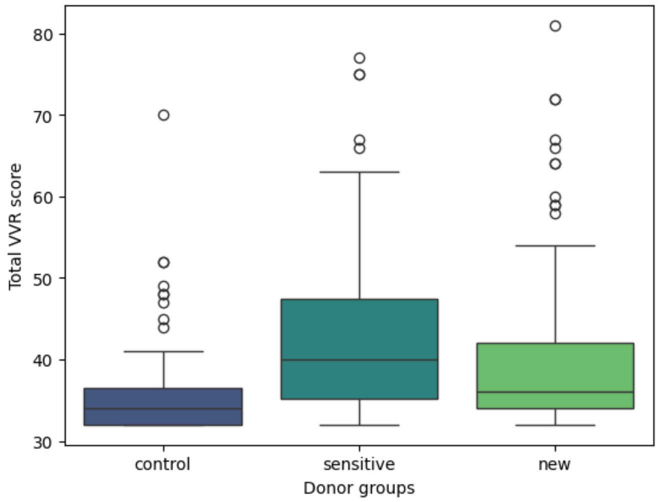



Figure 6: Boxplot representation of total VVR score by donor type

Analysis of Variances along with Tukey's Honestly Difference (Tukey's HSD) tests show that age (F(2)=1.34, p=0.26), body-mass index (F(2)=0.04, p=0.96), and total SSAS scores (F(2)=1.06, p=0.35) are not significantly

9 These findings align with previous literature which reports that women (Eder et al., 2008; Tomita et al., 2002), first-time donors (Ogata et al., 1980), and donors with unpleasant previous donation experiences are more likely to experience VVR (Ditto et al., 2012).

different between distinct donor types, while the total VVR scores are different ($F(2)=14.59$, $p<0.001$). The gap in the average VVR scores between the sensitive group and the rest is more pronounced (see Figure 6), while it is also significant for the control and first-time donors ($p=0.01$).

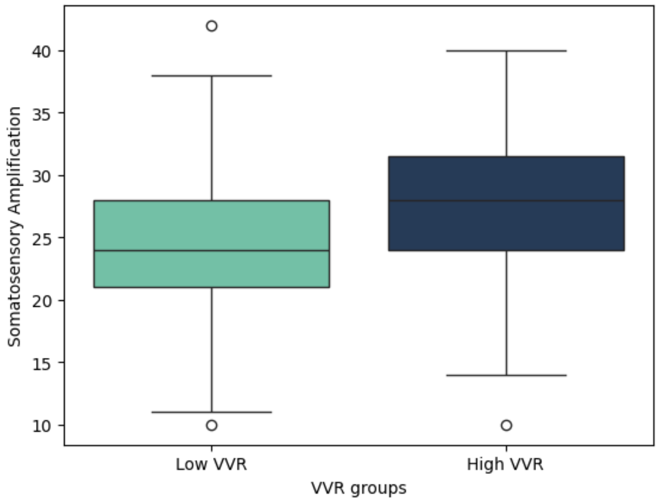



Figure 7: Boxplot representation of total SSAS score by VVR group

Figure 7 demonstrates a boxplot representation of the total SSAS score by the VVR group. The total SSAS score is relatively higher for the high VVR group, signaling an association between higher somatosensory amplification and the greater risk of VVR. A Mann-Whitney U test also confirmed that the distribution of total SSAS scores is significantly different between the two VVR groups (p-value<0.001). According to these findings, the information conveyed by the total SSAS score may contribute to distinguishing between the high and low VVR groups.

## 4.2 *Overview of Methodology and Evaluation Stages*

An overview of model implementation and performance evaluation is provided in Figure 8. XGBoost, MLP, and Random Forest models were trained independently on two datasets: the baseline dataset and the extended dataset. For each algorithm trained on the baseline dataset, two distinct pipelines were designed, incorporating ROS and SMOTE as the respective oversampling techniques. The results of nested cross-validation (outer loop) and test set were used to identify the best-performing models on the baseline dataset.

XGBoost, Random Forest, and MLP models with SMOTE were also trained on the extended dataset, which included an additional feature related to the somatosensory amplification. The performance of these models was compared across datasets to evaluate the impact of this additional feature. Then, the error analysis and feature importance were conducted using the best-performing model-dataset combination.

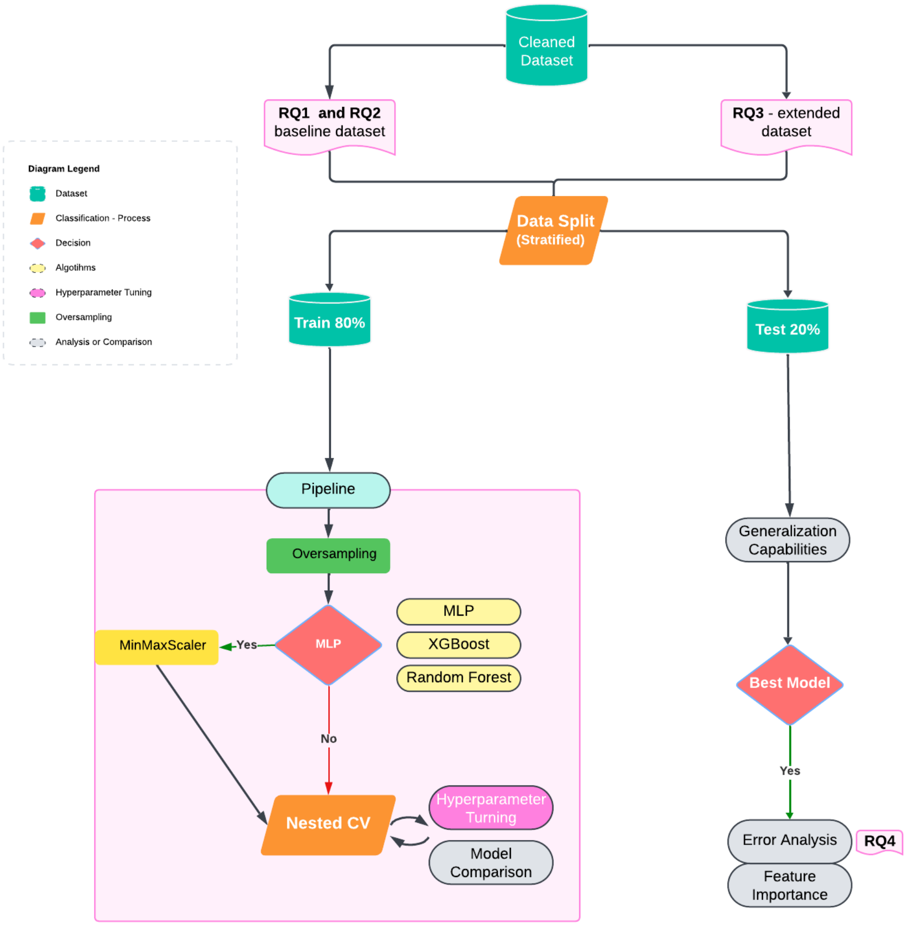



Figure 8: Overview of methodology and evaluation

## 4.3 *Algorithms*

### 4.3.1 *Baseline Models*

In this study, two baseline models, Support Vector Machine (SVM) (Cortes & Vapnik, 1995) and Dummy Classifier, were used to ensure a fair comparison of the model. SVM, employed by previous studies to predict anxiety (Gavrilescu & Vizireanu, 2019) and stress levels (Giannakakis et al., 2022), was used to quantify the relative prediction performance of the suggested algorithms. A Dummy Classifier was also provided to assess model performance compared to random guessing.

### 4.3.2 *Main Models*

MLP (Rosenblatt, 1958) was positioned as the main interest of this study considering its ability to capture non-linear complex relationships. As stated in section 2.4, several factors may jointly increase the risk of experiencing VVR and may interact with each other leading to non-linear relationships. Besides, the performance of MLP to utilize facial expressions when predicting VVR is well documented by previous studies (Rudokaite et al., 2024, 2023a). The MLP models used in these studies consistently outperformed the other models (XGBoost, Random Forest, and Decision Trees).

XGBoost (Chen & Guestrin, 2016) and Random Forest were selected, considering their proven performance on VVR predictions. XGBoost is stated as the best-performing model in Rudokaite et al. (2023b) with an F1 score of 0.86. Random Forest stands out as the second best-performing model in Rudokaite et al. (2023a) with an F1 score of 0.76. The implicit feature selection mechanism of Random Forest was also considered in this choice, given the moderate size of the feature set.

## 4.4 *Hyperparametertuning and Nested Cross Validation*

The cleaned dataset was split into two groups using stratified splits as shown in Figure 8; 80 percent as the training and 20 percent as the test set. The stratification process ensured that the class distribution in each split maintained the same proportions as the overall set. Nested cross-validation with GridSearchCV was employed for the hyperparameter selection and the evaluation of the models.

GridSearchCV was utilized to optimize the proposed models. This optimization framework was selected considering its characteristic of easy implementation, and enhanced reproducibility. While GridSearchCV systematically evaluates all possible combinations, it can be more resource-intensive than RandomSearch, especially for the high-dimensional hyperparameter spaces. Still, GridSearchCV was preferred over the RandomSearch, in this specific setup considering that it reduces the risk of missing promising combinations.

Nested cross-validation was preferred over other data-splitting methods by aiming to minimize the risk of data leakage and ensure more reliable evaluations. K-fold cross-validation and hold-out approaches tend to report more optimistic validation performances since they perform the hyperparameter selection and model evaluation using identical datasets. A 5-fold inner loop was used for hyperparameter selection. The average scores of a 10-fold outer loop were used for model comparison. The nested cross-validation was performed within pipelines applied GridSearchCV for each algorithm listed in Table 3.

Table 3: Hyperparameter space for GridSearchCV

| Algorithm | Parameter | Parameter Set |
|---|---|---|
| XGBoost | Max depth | [1, 2, 3] |
| | Subsample | [0.5, 0.7] |
| | Col sample by three | [0.5, 0.7] |
| | Learning rate | [0.001, 0.01] |
| | Alpha | [1, 5, 7, 10] |
| MLP | Hidden layer sizes | [(64,32), (64), (32), (16)] |
| | Activation function | [relu, tanh] |
| | Batch size | [32, 64] |
| | Alpha | [1, 2, 5, 10] |
| Random Forest | Max depth | [1, 2, 3, 5] |
| | Min samples split | [0.1, 0.2, 0.3] |
| | Min samples leaf | [2, 3, 5] |
| | Criterion | [gini, entropy] |
| | Number of estimators | [100, 200] |

### 4.5 *Evaluation Method and Feature Importance*

An evaluation set including precision, recall, accuracy, F1-score, AUC, and AUC-PR was estimated for inner and outer loops of the nested cross-validation along with the test set. All the metrics were used for the evaluation process, while the primary attention was on the F1 score. Recall score was also considered: the study prioritized minimizing the wrong classification of donors in the High VVR group, aiming to eliminate the serious medical consequences. However, models with a high recall score might tend to be biased towards a specific class, which can prevent models from learning effectively. Therefore, the predictive performance of the trained models was also compared with a dummy classifier to assess their performance compared to random guessing. AUC-PR was also utilized for the evaluations, considering the class imbalances in the sample. The class-specific evaluation metrics were reported for a more detailed comparison across models.

As described in section 4.2, the models were also trained on the extended dataset to investigate the SRQ3. The performance of the models was compared across baseline and extended datasets. This comparison enabled a fair assessment of the joint utilization of SSAS and facial action units in VVR predictions.

Finally, feature importance analysis was conducted with the best-performing model-dataset combination. Shapley Additive Explanation (SHAP) enabling both local and global interpretability, was preferred over the other methods.

### 4.6 *Error Analysis*

The best-performing model-dataset pair was used to produce confusion matrices by donor types. The analysis was also performed for the MLP model on the baseline dataset to document heterogeneities across models.

## 5 RESULTS

### 5.1 *Tuned Hyperparameters*

Table 4: Best parameters selected by GridSearchCV

| Algorithm | Parameter | Baseline Dataset | | Extended Dataset |
|---|---|---|---|---|
| | | SMOTE | ROS | SMOTE |
| XGBoost | Max depth | 3 | 1 | 1 |
| | Subsample | 0.5 | 0.5 | 0.5 |
| | Col sample by three | 0.7 | 0.7 | 0.5 |
| | Learning rate | 0.01 | 0.001 | 0.01 |
| | Alpha | 7 | 10 | 5 |
| MLP | Hidden layer sizes | (16) | (32) | (32) |
| | Activation function | tanh | relu | tanh |
| | Batch size | 32 | 32 | 32 |
| | Alpha | 1 | 1 | 5 |
| Random Forest | Max depth | 3 | 2 | 2 |
| | Min samples split | 0.3 | 0.3 | 0.3 |
| | Min samples leaf | 5 | 2 | 2 |
| | Criterion | gini | gini | gini |
| | Number of estimators | 100 | 100 | 100 |

Hyperparameter optimization was conducted using the GridSearchCV as described in section 4.4. Inner loop of the nested cross-validation was utilized to select the best-performing (highest average F1 score) hyperparameters for each algorithm. The tuned hyperparameters listed in Table 4 were employed in the analysis. Relatively shallow trees and single-layer neural networks were chosen over more complex models, suggesting that simple models might be sufficient to capture the existing relationships.

## 5.2 *Results of Nested Cross Validation*

The outer loop of the nested cross-validation is utilized for the initial comparison of the models' performances. Support Vector Machine and Dummy Classifier are provided to establish a baseline for this comparison.

Table 5: Model evaluation with nested cross validation

| Class Imbalance | Model | Recall | Precision | Accuracy | F1 | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|---|---|
| ROS | SVC | 0.42 (0.13) | 0.40 (0.10) | 0.59 (0.08) | 0.39 (0.08) | 0.63 (0.11) | 0.54 (0.08) |
| | XGBoost | 0.46 (0.15) | 0.42 (0.13) | 0.62 (0.08) | 0.43 (0.12) | 0.60 (0.10) | 0.55 (0.11) |
| | MLP | 0.88 (0.16) | 0.34 (0.04) | 0.39 (0.09) | 0.48 (0.04) | 0.54 (0.04) | 0.42 (0.06) |
| | Random Forest | 0.36 (0.17) | 0.37 (0.14) | 0.59 (0.08) | 0.35 (0.13) | 0.60 (0.11) | 0.51 (0.11) |
| SMOTE | SVC | 0.51 (0.19) | 0.42 (0.11) | 0.61 (0.09) | 0.45 (0.13) | 0.62 (0.10) | 0.51 (0.09) |
| | XGBoost | 0.49 (0.18) | 0.46 (0.19) | 0.62 (0.13) | 0.46 (0.16) | 0.63 (0.16) | 0.55 (0.16) |
| | MLP | 0.75 (0.20) | 0.35 (0.06) | 0.44 (0.13) | 0.46 (0.06) | 0.55 (0.08) | 0.44 (0.09) |
| | Random Forest | 0.48 (0.14) | 0.44 (0.13) | 0.62 (0.09) | 0.45 (0.12) | 0.61 (0.12) | 0.51 (0.10) |
| | Dummy classifier | 0.31 (0.09) | 0.34 (0.09) | 0.59 (0.06) | 0.32 (0.09) | 0.51 (0.07) | 0.34 (0.02) |

Among the algorithms paired with ROS as an oversampling technique, all models show improvement over the Dummy Classifier based on the F1 score. However, a closer examination of other metrics reveals contrasting findings. As shown in Table 5, MLP with ROS achieves the highest F1 score of 0.48, and a recall score of 0.88, however, it falls behind the dummy classifier based on accuracy, indicating a poorer performance than a random guess. MLP model accurately identifies donors at high risk of VVR, albeit at the cost of an increased rate of false positives. XGBoost with ROS stands out with a more balanced performance across precision and recall and reaches the highest accuracy (0.62) and PR-AUC score (0.55) among the models with ROS.

Implementing SMOTE as an oversampling technique particularly improves the overall performance of the Random Forest and SVC. All the metrics, except the PR-AUC score, indicate a better performance for the Random Forest when SMOTE is preferred over ROS (see Table 5). While

XGBoost (F1=0.46, Recall=0.49, Accuracy=0.62) remains the strongest model based on the outer loop of the nested cross-validation, Random Forest with SMOTE almost matches its performance with an F1 score of 0.45 and a recall of 0.48. Although the MLP model achieves the highest F1 score (0.46), it could not outperform the Dummy Classifier and SVC in terms of accuracy.

An overall assessment of the nested cross-validation results reported in Table 5 highlights three main findings. Models with SMOTE generally outperform models with ROS in most metrics. XGBoost and Random Forest stand out as the best-performing models with a more balanced performance than MLP. Although MLP models reach the highest F1 and recall scores, they substantially lag behind the baseline models in terms of accuracy.

## 5.3 *Generalization Abilities of Models*

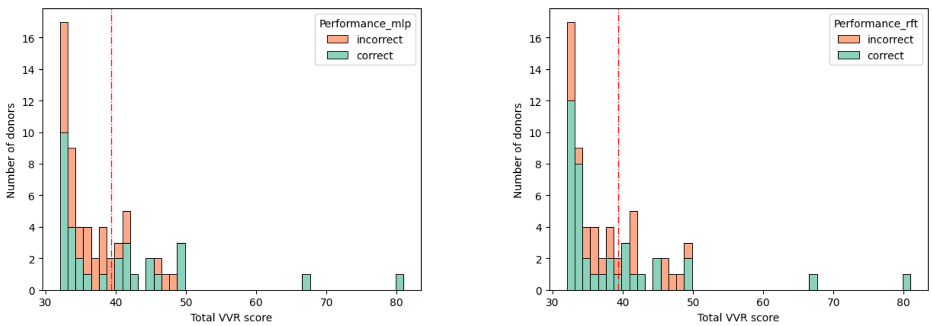
The results of the nested cross-validation were complemented with an assessment of the models in a hold-out unseen set. Table 6 reports the performances of the models in the test set, which was used to assess the generalization capabilities of the models.

Table 6: Model evaluation on test set

| Class Imbalance | Model | Recall | Precision | Accuracy | F1 | ROC-AUC | PR-AUC |
|---|---|---|---|---|---|---|---|
| ROS | SVC | 0.35 | 0.32 | 0.55 | 0.33 | 0.51 | 0.40 |
| | XGBoost | 0.45 | 0.30 | 0.48 | 0.36 | 0.45 | 0.30 |
| | MLP | 0.80 | 0.33 | 0.42 | 0.47 | 0.48 | 0.36 |
| | Random Forest | 0.40 | 0.33 | 0.55 | 0.36 | 0.48 | 0.35 |
| SMOTE | SVC | 0.55 | 0.32 | 0.48 | 0.41 | 0.52 | 0.38 |
| | XGBoost | 0.30 | 0.38 | 0.61 | 0.33 | 0.50 | 0.41 |
| | MLP | 0.70 | 0.37 | 0.52 | 0.48 | 0.50 | 0.32 |
| | Random Forest | 0.55 | 0.42 | 0.61 | 0.48 | 0.56 | 0.39 |
| | Dummy classifier | 0.40 | 0.42 | 0.63 | 0.41 | 0.57 | 0.36 |

As presented in Table 6, utilizing SMOTE improves all models' F1 scores, except XGBoost, where the improvement was more pronounced for Random Forest. The accuracy of the MLP model is also improved when SMOTE is used, increasing from 0.42 to 0.52. A comparison of PR-AUC scores reveals mixed results regarding the performance of the oversampling techniques. While the SVC and MLP models perform better with ROS in terms of PR-AUC scores, the tree-based models achieve higher scores with SMOTE. This suggests that the generalization performance of the tree-based models becomes more robust to threshold selection when SMOTE is used.

While the XGBoost with SMOTE shows promising performance in cross-validation (F1=0.46), its performance lags behind the baseline models (SVC and Dummy Classifier) in the test set, with an F1 score of 0.33. MLP and Random Forest models show consistent results across both nested cross-validation and test sets. Similar to the nested cross-validation, MLP models reach the highest recall score (ROS=0.80, SMOTE=0.70), while their poor accuracy and precision performance remain. As seen in Figure 9, the MLP model demonstrates lower performance in distinguishing between high and low VVR groups compared to the Random Forest. Random Forest and MLP with SMOTE achieve the highest F1 score (0.48), whereas Random Forest shows a more balanced performance across other metrics (Recall=0.55, Precision=0.42, Accuracy=0.61, PR-AUC=0.39). Although its performance based on accuracy and ROC-AUC is slightly lower than the Dummy Classifier, it outperforms the baseline models with a better Recall, PR-AUC, and F1 performance. To this end, Random Forest with SMOTE is identified as the best-performing model on the baseline dataset, through a joint evaluation of test and cross-validation performances.




(a) MLP with SMOTE

(b) Random Forest with SMOTE

Figure 9: Distribution of predictions categorized by accuracy

## 5.4 *Model Performance by Class*

The primary aim of this study is to identify donors with a high risk of VVR. Consequently, the evaluation metrics reported in earlier sections focus on the positive class. While minimizing false negatives is particularly important for this study, it should be complemented with capabilities to distinguish positive and negative classes. Models with high recall capabilities might tend to learn positive instances, even if this may result in a higher number of false positives. A class-specific comparison of evaluation metrics is particularly important considering these trade-offs.

Table 7: Generalization performance by class

| Class | Precision | Recall | F1-score | Support |
|---|---|---|---|---|
| | | ***Dummy Classifier*** | | |
| 0 | 0.72 | 0.74 | 0.73 | 42 |
| 1 | 0.42 | 0.40 | 0.41 | 20 |
| accuracy | | | 0.63 | 62 |
| macro avg. | 0.57 | 0.57 | 0.57 | 62 |
| weighted avg. | 0.62 | 0.63 | 0.63 | 62 |
| | | ***SVC*** | | |
| 0 | 0.68 | 0.45 | 0.54 | 42 |
| 1 | 0.32 | 0.55 | 0.41 | 20 |
| accuracy | | | 0.48 | 62 |
| macro avg. | 0.50 | 0.50 | 0.48 | 62 |
| weighted avg. | 0.56 | 0.48 | 0.50 | 62 |
| | | ***MLP*** | | |
| 0 | 0.75 | 0.43 | 0.55 | 42 |
| 1 | 0.37 | 0.70 | 0.48 | 20 |
| accuracy | | | 0.52 | 62 |
| macro avg. | 0.56 | 0.56 | 0.51 | 62 |
| weighted avg. | 0.63 | 0.52 | 0.53 | 62 |
| | | ***Random Forest*** | | |
| 0 | 0.75 | 0.64 | 0.69 | 42 |
| 1 | 0.42 | 0.55 | 0.48 | 20 |
| accuracy | | | 0.61 | 62 |
| macro avg. | 0.59 | 0.60 | 0.59 | 62 |
| weighted avg. | 0.64 | 0.61 | 0.62 | 62 |

Table 7 reports evaluation metrics of the selected models and baseline model by high and low VVR risk groups. Both Random Forest and MLP outperform the SVC model for every class. MLP and Random Forest models reach the same F1 score for the high-risk group (0.48), while the Random Forest outperforms the MLP based on the low-risk group with an F1 score of 0.69. MLP model accurately identifies a higher number of donors with high VVR risk, while it tends to misclassify donors with low VVR risk under the high-risk group.

A comparison of models with the Dummy Classifier provides further insights regarding the limitations of the models. MLP exhibits a lower recall performance for the negative class, and lower precision performance for the positive class compared to the Dummy Classifier. Random Forest also could not outperform the Dummy Classifier based on the low-risk group, while it still stands out as the best performer since the high-risk group is the primary interest of this study.

Figure 10 demonstrates the confusion matrices of the models. MLP shows the highest recall performance for the high-risk group, the XGBoost achieves the highest recall for the low-risk group, and the Random Forest demonstrates a more balanced performance compared to these two models.

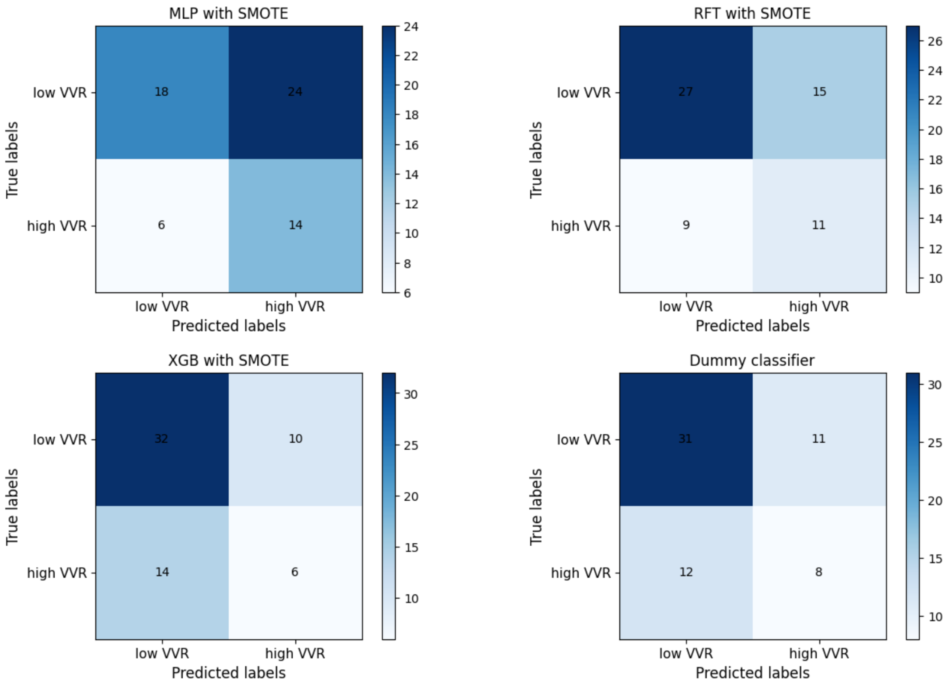



Figure 10: Confusion matrices of the selected models

## 5.5 *Model Performance on Extended Dataset*

The models were also independently trained on the extended dataset containing the total SSAS score along with the time characteristics of facial action units. Table 8 reports performance of the models on both the baseline and extended datasets for a straightforward comparison.

The Random Forest model shows similar performances across the datasets. The F1 score based on the test set slightly improves when somatosensory amplification is introduced (Recall=0.55, Precision=0.44, Accuracy=0.63, F1=0.49, PR-AUC=0.38). The cross-validation performance of XGBoost models shows no significant difference across the datasets, while the accuracy and precision performances on the test set improve marginally with the extended dataset (Recall=0.35, Precision=0.47, Accuracy=0.66, F1=0.40, PR-AUC=0.38). Conversely, the MLP model demonstrates better performance on the baseline dataset based on test set performance (Recall=0.70, Precision=0.37, Accuracy=0.52, F1=0.48, PR-AUC=0.32). The performance gap between the nested cross-validation and test sets signals overfitting and poor generalization capabilities on the extended dataset for the MLP model.

Table 8: Model performance across datasets

| Model | Dataset | | Recall | Precision | Accuracy | F1 | ROC-AUC | AUC-PR |
|---|---|---|---|---|---|---|---|---|
| **Random Forest** | Baseline | Nested CV | 0.48 | 0.44 | 0.62 | 0.45 | 0.61 | 0.51 |
| | | Test | 0.55 | 0.42 | 0.61 | 0.48 | 0.56 | 0.39 |
| | Extended | Nested CV | 0.48 | 0.42 | 0.62 | 0.44 | 0.61 | 0.53 |
| | | Test | 0.55 | 0.44 | 0.63 | **0.49** | 0.55 | 0.38 |
| **MLP** | Baseline | Nested CV | 0.75 | 0.35 | 0.44 | 0.46 | 0.55 | 0.44 |
| | | Test | 0.70 | 0.37 | 0.52 | 0.48 | 0.50 | 0.32 |
| | Extended | Nested CV | 0.61 | 0.36 | 0.52 | 0.43 | 0.58 | 0.46 |
| | | Test | 0.70 | 0.31 | 0.40 | 0.43 | 0.38 | 0.28 |
| **XGBoost** | Baseline | Nested CV | 0.49 | 0.46 | 0.62 | 0.46 | 0.63 | 0.55 |
| | | Test | 0.30 | 0.38 | 0.61 | 0.33 | 0.50 | 0.41 |
| | Extended | Nested CV | 0.46 | 0.46 | 0.63 | 0.45 | 0.62 | 0.55 |
| | | Test | 0.35 | 0.47 | 0.66 | 0.40 | 0.52 | 0.38 |
| **Dummy classifier** | | | 0.40 | 0.42 | 0.63 | 0.41 | 0.57 | 0.36 |

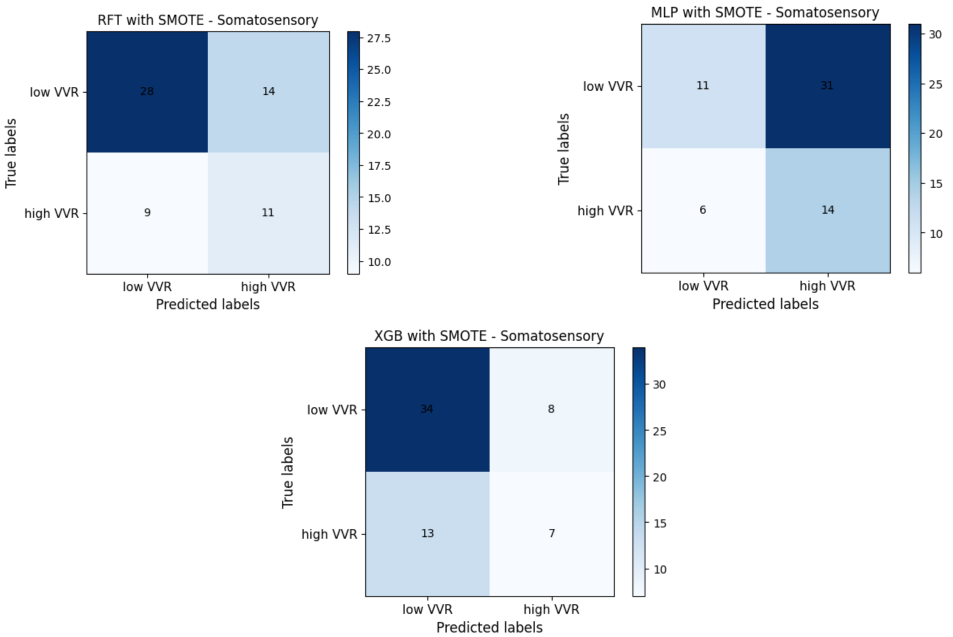
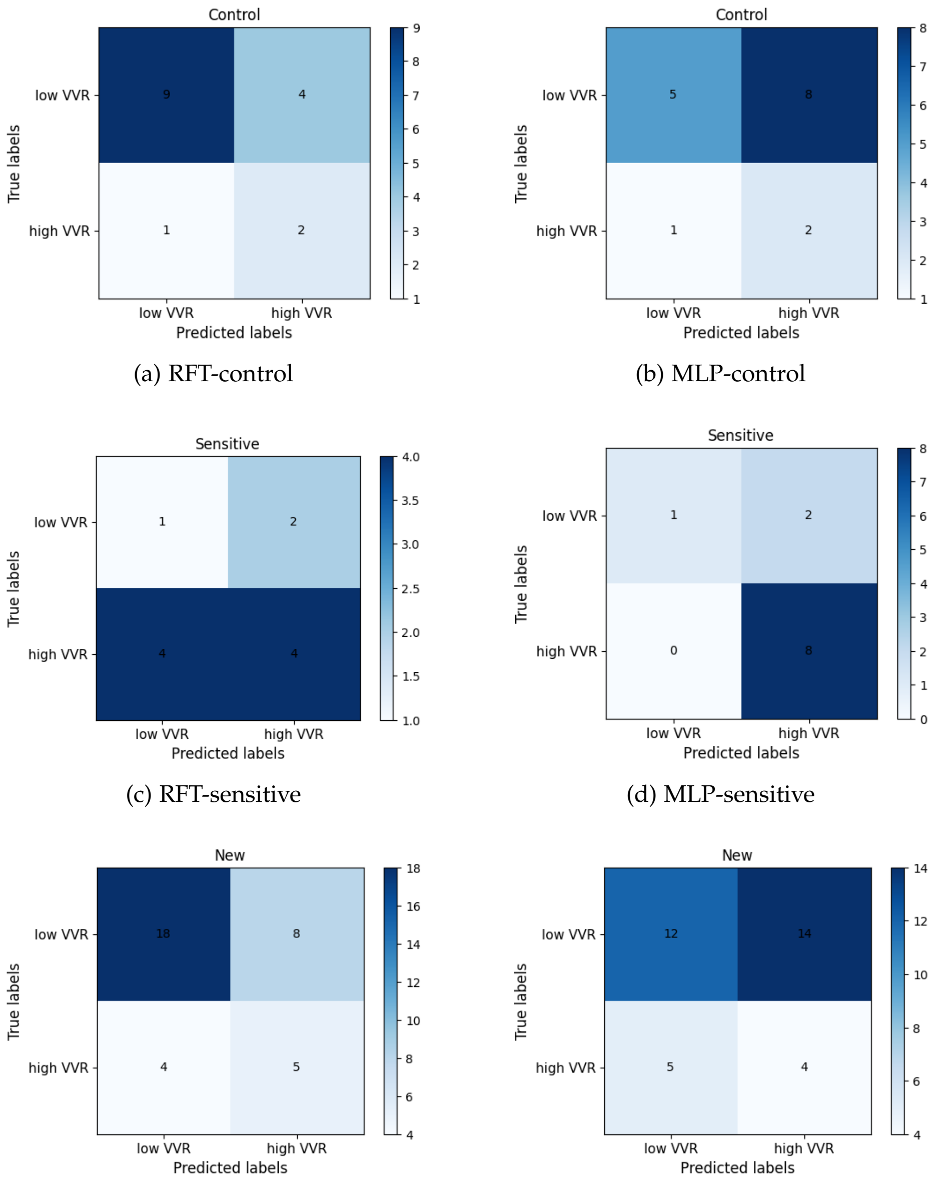



Figure 11: Confusion matrices of the models on the extended dataset

A comparison of Figure 10 and Figure 11 shows that incorporating somatosensory amplification does not change the number of correctly classified positive instances for the MLP and Random Forest models. While the classification performance for low-risk group slightly improves with the Random Forest, it decreases for the MLP. XGBoost shows slight improvement for each class when SSAS score included. Overall, Random Forest model demonstrates a more balanced performance across classes compared to the MLP and XGBoost.

## 5.6 *Error Analysis*




(a) RFT-control

(b) MLP-control

(c) RFT-sensitive

(d) MLP-sensitive

(e) RFT-first time donors

(f) MLP-first time donors

Figure 12: Confusion matrices of best models by donor type

The error analysis was conducted using the best-performing model, Random Forest with SMOTE. The MLP model is provided for comparison purposes.

Figure 12 provides significant insights regarding the learning capabilities of the models across donor types. The Random Forest model (RFT) achieves the highest recall rate for the control group within the positive class, while the recall performances for control and first-time donors remain similar within the negative class. The poorer recall performance was recorded for the sensitive group both in the positive and negative classes. Conversely, the MLP model demonstrates the highest recall performance for the sensitive group, where it correctly classified all positive instances. However, the sensitive group also shows poorer recall performance within the negative class where one of the two true negatives was identified as a positive instance.

Within-group comparison across models reveals the heterogeneity in the relative performances. While the MLP outperformed the Random Forest for the sensitive group, the Random Forest model performs better in classifying first-time donors and the control group into high and low risk categories.

### 5.7 *Feature Importance*

Feature importance analysis was conducted using the Random Forest Model on the extended dataset. Figure 13 reports the SHAP values of the most important 15 features. As demonstrated by the figure, mean changes in the intensity of lip tightening (AU23) and jaw-dropping (AU26) are identified as the most important features [10]. Mean changes correspond to the average change in facial action unit intensities between two consecutive time points. Since it captures the sudden changes in the intensity of facial action units, may serve as a good proxy for capturing unintentional facial expressions. The SHAP values show that a higher value of these mean changes increases the likelihood of classifying a donor as a high VVR group.

Characteristics related to blink (AU45) and lips part (AU25) also stand out as important features. Likewise, a higher value of these features is associated with a high VVR group. On the other hand, a higher median level of cheek raising (AU6) and lip corner depressor (AU15) increases the likelihood of classifying instances as low VVR group.

[10] A more detailed discussion on action units is provided in section 6.1

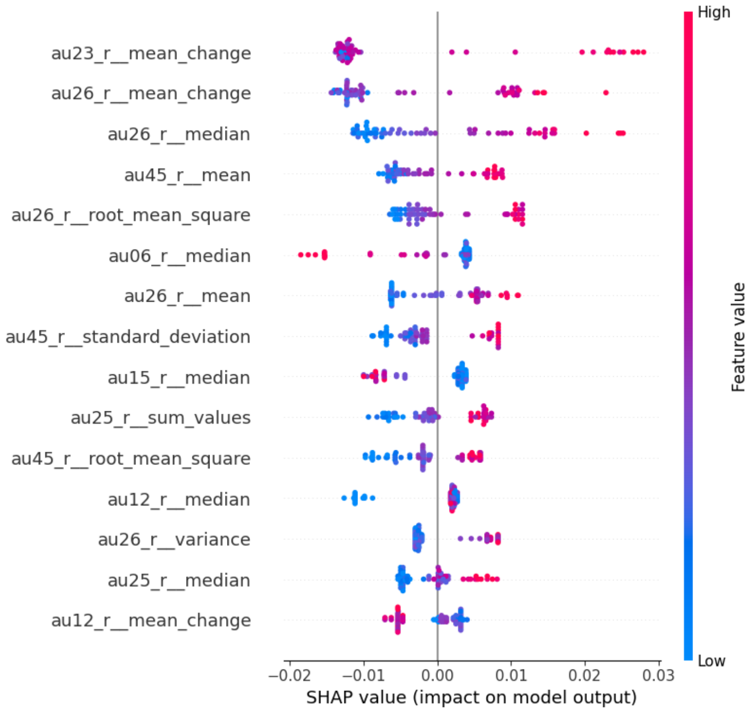



Figure 13: Feature contribution to model performance (Random Forest on the extended dataset)

# 6 DISCUSSION

## 6.1 *Summary and Interpretation of Results*

**SRQ1:** Overall review of the test and the nested cross-validation results on the baseline dataset shows that the Random Forest with SMOTE outperforms the MLP and XGBoost models by reaching the highest F1 score (0.48) while ensuring a more balanced performance (Accuracy=0.61, Precision=0.42, Recall=0.55). Although MLP with SMOTE model attained the same F1 score on the test set and a higher recall performance (0.70), its precision and accuracy performance lag behind the dummy classifier, indicating poorer generalization capabilities.

Rudokaite et al. (2023a) identified the MLP as the best-performing model, achieving the highest F1 and recall performance (Precision=0.79, Recall=0.84, F1=0.82, PR-AUC=0.79). The Random Forest (Precision=0.72, Recall=0.81, F1=0.76, PR-AUC=0.72) also stands out as the second-best performing model based on their results. While the Random Forest performs well in line with the previous literature, still lagging behind the state-of-art, the results of our study show contrasting results for MLP. MLP models underperformed in this study compared to their counterparts, particularly in terms of precision and accuracy. While these contrasting results might be partially attributed to the limitation of the study, will be discussed shortly, differences in the datasets also play a substantial role.

Both Rudokaite et al. (2023a) and this study utilize the data collected by the Faint Project (FAINT, 2018), while there exist significant differences between the respective datasets. Rudokaite et al. (2023a) utilizes data on facial action unit intensities only from the first two stages of donation. These intensity levels were extracted from video recordings lasting from 1 to 2 minutes for each stage. Although our study utilized data from the first two stages, most of the extracted information comes from stage 4, when donors are seated on the donation chair and awaiting needle insertion. The facial expressions observed during the earlier stages of donation might provide more valuable insights than those from later stages, and this could account for the variability in model performances across studies. In particular, if the high VVR group exhibits earlier responses triggered by psychological stimuli before the needle insertion phase, including data from stage 4 may be redundant and pose a risk for noisier estimates. The responses to psychological stimuli such as anticipated pain, fear, and anxiety prior to donation, might be better predictors of VVR than responses to physical stimuli like seeing a needle or waiting at the donation chair. These results align with a part of previous literature that reports elevated physiological stress response (systolic and diastolic blood pressure, baroreflex sensitivity) prior to donation (Kaur et al., 2023).[11]

**SRQ2:** Models with SMOTE outperformed the models with ROS in test sets based on F1 scores, except XGBoost. SMOTE also enhanced the precision and accuracy performance of each model. However, comparing models based on PR-AUC scores reveals contrasting results, while tree-based models demonstrate better performance on the test set when SMOTE is utilized, the PR-AUC scores decline for the rest with the implementation of SMOTE. An overall review of the results implies that the ROS could not improve the prediction performance of the models compared to SMOTE. Despite the SMOTE being a more resource-intensive technique, it may still be valuable to implement, particularly for the three-based models.

**SRQ3:** A comparison of the selected models on the baseline and extended datasets shows that introducing somatosensory amplification to the feature set provides a limited improvement in XGBoost and Random Forest, while its overall effects remain model-dependent. The F1 score of the Random Forest slightly improves on the test set (baseline=0.48,

[11] The literature does not reach a consensus on the physiological stress response during a blood donation. There are also studies report that the physiological stress response increases up until a needle insertion and decreases afterward (Hoogerwerf et al., 2018).

extended=0.49), a similar improvement is not recorded in the results of the nested cross-validation. XGBoost shows a moderate improvement, performing better across all metrics except, for PR-AUC scores, its performance on nested cross-validation remains similar to that of the baseline dataset. Even though the three-based models show limited improvement with the inclusion of somatosensory amplification, the MLP model shows better performance on the baseline dataset.

Feature importance analysis conducted with the Random Forest quantified the limited effect of the somatosensory amplification. Despite its ranking as the 30th most important feature (out of 154), it is not among the primary drivers of the prediction. These findings do not rule out the contribution of somatosensory amplification to the VVR predictions, still, it raises questions on its joint utilization with the facial action unit intensities. If the baseline dataset already captures aspects of somatosensory amplification (as being a proxy of somatosensory amplification), then incorporating a stand-alone feature might introduce noise to the predictions. The poorer generalization capacity of the MLP model on the extended dataset might partially be explained by this, considering that neural networks are less prone to noisy data and uninformative features compared to tree-based models (Grinsztajn et al., 2022).

Still, the feature importance analysis on the extended dataset provides valuable insights into the dynamics of VVR predictions. The mean change in lip tightening (AU23) and jaw-dropping (AU26) are revealed as the most important predictors of the high VVR group, while the higher median level of cheek-raising (AU06) is associated with a low VVR group. Several time characteristics of blinking (AU45) also emerge as important predictors for the positive class. These results are aligned with existing literature: the lip tightener (AU23) and jaw-dropping (AU26) are linked to a greater heart rate reactivity (Blasberg et al., 2023), a well-known stress marker. The spontaneous blink rate is also indicated in the literature as one of the markers of detecting underlying pain (Paparella et al., 2020). In contrast, cheek raiser (AU06) is reported as an indicator of a genuine smile, a signal of positive feelings (Kawulok et al., 2021).

**SRQ4:** The error analysis conducted at the last stage of the model implementation phase outlines the significant heterogeneity in prediction performances by donor groups. Random Forest demonstrates the best performance in classifying the first-time donors and control group, while the worst performance is recorded for the sensitive group.

## 6.2 *Limitations*

The implementation of the methodology includes several decisions that bring their unique challenges. The main limitation of the study is the categorization of the high and low VVR groups. The mean VVR score is used as a threshold for class categorization. This selection is made considering that the mean values incorporate information from all the available data points, which do not solely focus on the donors from the two extremes of the VVR distribution. However, considering right- skewed distribution of the total VVR scores, the mean values may introduce a counterfactual threshold which may not accurately represent the actual data, while also not prone to the outliers. Such an approach may represent inherently diverse donors as similar.

Another limitation is the lack of a stand-alone feature selection process. The feature set of this study shows high-dimensional characteristics considering the small dataset in hand. While the L1 and L2 regularization techniques have been introduced to the analysis as a parameter of the algorithms, these techniques might not completely mitigate the noise posed by the uninformative features. Still, this should not be a substantial caveat considering that the previous literature reports a limited contribution in the existence of recursive feature elimination (Rudokaite et al., 2023a; Rudokaite et al., 2023b).

One notable drawback is related to the hyperparameter selection process. GridSearchCV is one of the most common methods that enhance interpretability and flexibility, however, the whole optimization process relies on the predefined hyperparameter space. Defining this hyperparameter space might be challenging and including a higher set of parameters is computationally expensive.

## 6.3 *Recommendations for Future Work*

The limitations discussed in section 6.2 also highlight the areas where the study can be further improved. Employing alternative methods to define the VVR classes should be on the agenda of future studies. Utilizing alternative methods to define classification thresholds (e.g., median levels which are prone to outliers) might better represent the actual VVR distribution. Alternatively, further studies can be designed to predict continuous VVR risk scores, which may alleviate the risk of introducing counterfactual thresholds.

Future studies may also aim to gain deeper insights from facial expressions. A joint presence of the facial action units, which may serve as proxies of underlined emotions may improve the exploration of complex relationships between the facial action units. Focusing on the most informative predictors identified in this study, such as blinking (AU45) and action units around the lips (AU23-AU25), could also enhance VVR predictions.

Alternative approaches (e.g., Cranbach's alpha test and PCA) to better utilize information derived from the somatosensory amplification and investigating its stand-alone effect in VVR predictions might be another noteworthy attempt for future studies.

While the suggestions above mostly focus on research design, some improvements can also made in the practical implementation. Indeed, technical aspects of model implementation, such as documenting the performance of class weights over standalone oversampling techniques, or exploring the effects of alternative hyperparameter optimization methods (e.g Bayesian optimization) might be documented in future studies.

### 6.4 *Contribution to Society*

While the previous literature mostly relies on observable and self-reported characteristics (such as age, gender, pre-donation history, etc.) to predict VVR, the projects conducted under the Faint project show promising results that the machine learning algorithms and improvements in video recording technology may extend the scope of the observable and unobservable characteristics in VVR predictions. These studies successfully utilize facial micro-expressions and thermal images, enabling auto-detection mechanisms to predict VVR. The Faint Project (FAINT, 2018) is the main inspiration for this thesis aiming to predict the risk of experiencing VVR by exploiting data on blood donors's facial micro-expressions. Despite the models utilized in this project did not outperform the current state-of-the-art (Rudokaite et al., 2023a), the results of the study ensure a robust foundation and clear guidance for advancing future studies on the utilization of facial expressions in VVR predictions and auto-detection mechanisms. Facial expression, instrumental in identifying the underlying emotions, can enhance not only VVR predictions but also have a potential for diverse applications ranging from mental health assessments to security.

## 7 CONCLUSION

This study utilizes facial micro-expressions from an earlier stage of blood donation, prior to the needle insertion, aiming to classify donors based on their likelihood to experience low and high levels of VVR during later stages of donation. It aims to address the main research question, outlined by several sub-questions (see section 2.4). The results of the study reveal that Random Forest outperforms the MLP and XGBoost with a more balanced and robust performance. This finding aligns closely with the broader literature (Caruana & Niculescu-Mizil, 2006; Fernández-Delgado et al., 2014), demonstrating the superior performance of tree-based models compared to Neural Networks, especially on the tabular data (Grinsztajn et al., 2022). Incorporating SMOTE as an oversampling technique improves models' generalization capabilities compared to ROS, while the enhancement is more prominent for tree-based models. The study also demonstrates that the joint utilization of somatosensory amplification and facial action unit intensities has a limited contribution to the prediction performance of only tree-based models, with its effect being model-dependent. Despite a limited improvement, this finding raises concerns regarding the joint use of facial action units with somatosensory amplification. While the models trained within the scope of this study do not outperform the current state-of-the-art (Rudokaite et al., 2023a), the results extensively discussed in section 6 provide valuable insights and may serve as a foundation for future studies.

## REFERENCES

Almutairi, H., Salam, M., Alajlan, A., Wani, F., Al-Shammari, B., & Al-Surimi, K. (2017). Incidence, predictors and severity of adverse events among whole blood donors. *PLoS One*, *12*(7), e0179831. https://doi.org/10.1371/journal.pone.0179831

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018). Openface 2.0: Facial behavior analysis toolkit. *Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66.

Barends, H., Claassen-van Dessel, N., van der Wouden, J. C., Twisk, J. W. R., Terluin, B., van der Horst, H. E., & Dekker, J. (2020). Impact of symptom focusing and somatosensory amplification on persistent physical symptoms: A three-year follow-up study [Epub 2020 May 8]. *Journal of Psychosomatic Research*, *135*, 110131. https://doi.org/10.1016/j.jpsychores.2020.110131

Bargshady, G., Zhou, X., Deo, R. C., Soar, J., Whittaker, F., & Wang, H. (2020). Enhanced deep learning algorithm development to detect pain intensity from facial expression images. *Expert Systems with Applications*, *149*, 113305. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113305

Barsky, A. J. (1979). Patients who amplify bodily sensations. *Annals of Internal Medicine*, *91*(1), 63–70. https://doi.org/10.7326/0003-4819-91-1-63

Barsky, A. J., Goodson, J. D., Lane, R. S., & Cleary, P. D. (1988). The amplification of somatic symptoms. *Psychosomatic Medicine*, *50*(5), 510–519. https://doi.org/10.1097/00006842-198809000-00007

Barsky, A. J., & Silbersweig, D. A. (2023). The amplification of symptoms in the medically ill [Epub 2022 Jul 12]. *Journal of General Internal Medicine*, *38*(1), 195–202. https://doi.org/10.1007/s11606-022-07699-8

Barsky, A. J., Wyshak, G., & Klerman, G. L. (1990). The somatosensory amplification scale and its relationship to hypochondriasis. *Journal of Psychiatric Research*, *24*(4), 323–334. https://doi.org/10.1016/0022-3956(90)90004-a

Blasberg, J. U., Gallistl, M., Degering, M., Baierlein, F., & Engert, V. (2023). You look stressed: A pilot study on facial action unit activity in the context of psychosocial stress. *Comprehensive Psychoneuroendocrinology*, *15*, 100187. https://doi.org/https://doi.org/10.1016/j.cpnec.2023.100187

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning*, 161–168. https://doi.org/10.1145/1143844.1143865

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, *abs/1603.02754*. http://arxiv.org/abs/1603.02754

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh—a python package). *Neurocomputing*, *307*, 72–77. https://doi.org/10.1016/j.neucom.2018.03.067

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Ditto, B., France, C. R., & Holly, C. (2010). Applied tension may help retain donors who are ambivalent about needles. *Vox Sanguinis*, *98*(3 Pt 1), e225–e230. https://doi.org/10.1111/j.1423-0410.2009.01273.x

Ditto, B., Gilchrist, P. T., & Holly, C. D. (2012). Fear-related predictors of vasovagal symptoms during blood donation: It's in the blood. *J Behav Med*, *35*(4), 393–399. https://doi.org/10.1007/s10865-011-9366-0

Eder, A. F., Hillyer, C. D., Dy, B. A., Notari, E. 4., & Benjamin, R. J. (2008). Adverse reactions to allogeneic whole blood donation by 16- and 17-year-olds. *JAMA*, *299*(19), 2279–2286. https://doi.org/10.1001/jama.299.19.2279

Ekman, P., & Friesen, W. V. (1976). Measuring facial movement. *Environmental Psychology & Nonverbal Behavior*, *1*(1), 56–75. https://doi.org/10.1007/BF01115465

FAINT. (2018). Faint: Facial infrared thermal imaging in een serious smartphone game [Accessed: 2023-21-09]. https://research.tilburguniversity.edu/en/projects/faint-facial-infrared-thermal-imaging-in-een-serious-smartphone-g

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*(90), 3133–3181. http://jmlr.org/papers/v15/delgado14a.html

France, C. R., France, J. L., Conatser, R., Lux, P., McCullough, J., & Erickson, Y. (2019). Predonation fears identify young donors at risk for vasovagal reactions [Epub 2019 Jul 2]. *Transfusion*, *59*(9), 2870–2875. https://doi.org/10.1111/trf.15424

France, C. R., France, J. L., Wissel, M. E., Ditto, B., Dickert, T., & Himawan, L. K. (2013). Donor anxiety, needle pain, and syncopal reactions combine to determine retention: A path analysis of two-year donor return data. *Transfusion*, *53*(9), 1992–2000. https://doi.org/10.1111/trf.12069

France, C. R., Rader, A., & Carlson, B. (2005). Donors who react may not come back: Analysis of repeat donation as a function of phlebotomist ratings of vasovagal reactions. *Transfusion and Apheresis Science*, *33*(2), 99–106. https://doi.org/10.1016/j.transci.2005.02.005

Garozzo, G., Crocco, I., Giussani, B., Martinucci, A., Monacelli, S., & Randi, V. (2010). Adverse reactions to blood donations: The read project. *Blood Transfusion*, *8*, 49–62. https://doi.org/10.2450/2009.0089-09

Gavrilescu, M., & Vizireanu, N. (2019). Predicting depression, anxiety, and stress levels from videos using the facial action coding system. *Sensors*, *19*(17). https://doi.org/10.3390/s19173693

Giannakakis, G., Koujan, M. R., Roussos, A., et al. (2022). Automatic stress analysis from facial videos based on deep facial action units recognition. *Pattern Analysis and Applications*, *25*, 521–535. https://doi.org/10.1007/s10044-021-01012-9

Gilchrist, P. T., McGovern, G. E., Bekkouche, N., Bacon, S. L., & Ditto, B. (2015). The vasovagal response during confrontation with blood-injury-injection stimuli: The role of perceived control. *Journal of Anxiety Disorders*, *31*, 43–48. https://doi.org/https://doi.org/10.1016/j.janxdis.2015.01.009

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [http://www.deeplearningbook.org]. MIT Press.

Google, I. (2024). Google colab. https://colab.research.google.com/

Grammarly, I. (2024). Grammarly. https://www.grammarly.com/

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? https://arxiv.org/abs/2207.08815

Hamilton, J. G. (1995). Needle phobia: A neglected diagnosis. *Journal of Family Practice*, *41*(2), 169–175.

Hammal, Z., & Cohn, J. F. (2012). Automatic detection of pain intensity. *Proceedings of the 14th ACM international conference on Multimodal interaction*. https://doi.org/10.1145/2388676.2388688

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with numpy. *Nature*, *585*, 357–362.

Hoogerwerf, M. D., Veldhuizen, I. J. T., Tarvainen, M. P., Merz, E.-M., Huis in 't Veld, E. M. J., de Kort, W. L. A. M., Sluiter, J. K., & Frings-Dresen, M. H. W. (2018). Physiological stress response patterns during a blood donation. *Vox Sanguinis*, *113*(4), 362–369. https://doi.org/10.1111/vox.12646

Huis In 't Veld, E. M. J., de Kort, W. L. A. M., & Merz, E. M. (2019). Determinants of blood donation willingness in the european union: A cross-country perspective on perceived transfusion safety, concerns, and incentives [Epub 2019 Feb 27]. *Transfusion*, *59*(4), 1273–1282. https://doi.org/10.1111/trf.15209

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Jones, E., et al. (2007). Scipy: Open source scientific tools for python. *Computing in Science & Engineering*, *9*(3), 10–20. https://doi.org/10.1109/MCSE.2007.58

Junder, A., Sánchez, J., & Hüllermeier, G. (2018). Imbalanced-learn: A python toolbox for machine learning with imbalanced data. *Journal of Machine Learning Research*, *18*(1), 1–8. https://imbalanced-learn.org/

Kaur, A., Kaur, R., Sood, T., Malhotra, A., Arun, P., Mittal, K., Kaur, P., Kaur, G., & Prakash, K. (2023). Physiological and psychological stress response of blood donors during the blood donation process. *Vox Sanguinis*, *118*(12), 1061–1068. https://doi.org/https://doi.org/10.1111/vox.13541

Kawulok, M., Nalepa, J., Kawulok, J., & Smolka, B. (2021). Dynamics of facial actions for assessing smile genuineness. *PLOS ONE*, *16*(1), e0244647. https://doi.org/10.1371/journal.pone.0244647

Köteles, F., & Witthöft, M. (2017). Somatosensory amplification – an old construct from a new perspective. *Journal of Psychosomatic Research*, *101*, 1–9. https://doi.org/https://doi.org/10.1016/j.jpsychores.2017.07.011

Kumari, S. (2015). Prevalence of acute adverse reactions among whole blood donors: A 7 years study. *Journal of Applied Hematology*, *6*, 148–153.

Lucid Software, I. (2024). Lucidchart. https://www.lucidchart.com/

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774. https://shap.readthedocs.io/

McKinney, W. (2010). *Pandas: A foundational python library for data analysis*. https://pandas.pydata.org/

Meade, M. A., France, C. R., & Peterson, L. M. (1996). Predicting vasovagal reactions in volunteer blood donors. *Journal of Psychosomatic Research*, *40*(5), 495–501. https://doi.org/https://doi.org/10.1016/0022-3999(95)00639-7

Mikulski, B. (2023). Missingno: Missing data visualization module for python. https://github.com/ResidentMario/missingno

Newman, B. H., Newman, D. T., Ahmad, R., & Roth, A. J. (2006). The effect of whole-blood donor adverse events on blood donor return rates. *Transfusion*, *46*(8), 1374–1379. https://doi.org/10.1111/j.1537-2995.2006.00905.x

Ogata, H., Iinuma, N., Nagashima, K., & Akabane, T. (1980). Vasovagal reactions in blood donors. *Transfusion*, *20*(6), 679–683. https://doi.org/10.1046/j.1537-2995.1980.20681057157.x

Olatunji, B. O., Etzel, E. N., & Ciesielski, B. G. (2010). Vasovagal syncope and blood donor return: Examination of the role of experience and affective expectancies. *Behavior Modification*, *34*(2), 164–174. https://doi.org/10.1177/0145445510362576

OpenAI. (2024). Chatgpt (december 2024 version) [[Large language model]]. https://chat.openai.com

Paparella, M., Di Stasi, V. L., Mancini, F., De Rossi, D., Casagrande, M., Giannini, G., Cattaneo, L., & Montagnese, S. (2020). Painful stimulation increases spontaneous blink rate in healthy subjects. *Scientific Reports*, *10*(1), 20647. https://doi.org/10.1038/s41598-020-76804-w

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://scikit-learn.org/

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rudokaite, J., Ertugrul, I., Ong, L., Janssen, M., & Huis In't Veld, E. (2024). Face the needle: Predicting risk of fear and fainting during blood donation through video analysis. *2024 IEEE 18th International Con-*

*ference on Automatic Face and Gesture Recognition, FG 2024*. https://doi.org/10.1109/FG59268.2024.10581999

Rudokaite, J., Ertugrul, I. O., Ong, S., Janssen, M. P., & Huis in 't Veld, E. (2023a). Predicting vasovagal reactions to needles from facial action units. *Journal of Clinical Medicine*, *12*(4). https://doi.org/10.3390/jcm12041644

Rudokaite, J., Ong, L. S., Onal Ertugrul, I., Janssen, M. P., & Huis In 't Veld, E. M. J. (2023b). Predicting vasovagal reactions to needles with anticipatory facial temperature profiles. *Scientific Reports*, *13*(1), 9667. https://doi.org/10.1038/s41598-023-36207-z

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, 57–61. https://www.statsmodels.org/

Sokolowski, C. J., Giovannitti, J. A. J., & Boynes, S. G. (2010). Needle phobia: Etiology, adverse consequences, and patient management. *Dental Clinics of North America*, *54*(4), 731–744. https://doi.org/10.1016/j.cden.2010.06.012

Thijsen, A., & Masser, B. (2019). Vasovagal reactions in blood donors: Risks, prevention and management [Suppl 1]. *Transfusion Medicine*, *29*, 13–22. https://doi.org/10.1111/tme.12488

Thijsen, A., Masser, B., Gemelli, C. N., & Davison, T. E. (2019). Trends in return behavior after an adverse event in australian whole blood and plasma donors [First published: 12 August 2019, Citations: 23]. *Transfusion*, *59*(10). https://doi.org/10.1111/trf.15475

Thomson, R. A., Bethel, J., Lo, A. Y., Ownby, H. E., Nass, C. C., & Williams, A. E. (1998). Retention of "safe" blood donors. the retrovirus epidemiology donor study. *Transfusion*, *38*(4), 359–367. https://doi.org/10.1046/j.1537-2995.1998.38498257374.x

Tomita, T., Takayanagi, M., Kiwada, K., Mieda, A., Takahashi, C., & Hata, T. (2002). Vasovagal reactions in apheresis donors. *Transfusion*, *42*(12), 1561–1566. https://doi.org/10.1046/j.1537-2995.2002.00241.x

Vallat, R. (2018). Pingouin: Statistics in python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026

Viegas, C., Lau, S.-h., Maxion, R. A., & Hauptmann, A. (2018). Towards independent stress detection: A dependent model using facial action units. *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, 1–6. https://api.semanticscholar.org/CorpusID:53223721

Wang, H.-H., Chen, P.-M., Lin, C.-L., Jau, R.-C., Hsiao, S.-M., & Ko, J.-L. (2019). Joint effects of risk factors on adverse events associated with adult blood donations. *Medicine*, *98*(44), e17758. https://doi.org/10.1097/MD.0000000000017758

Waskom, M. L. (2021). *Seaborn: Statistical data visualization* (No. *60*). https://doi.org/10.21105/joss.03021

Werner, P., Al-Hamadi, A., Niese, R., Walter, S., Gruss, S., & Traue, H. C. (2014). Automatic pain recognition from video and biomedical signals. *2014 22nd International Conference on Pattern Recognition*, 4582–4587. https://doi.org/10.1109/ICPR.2014.784

World Health Organization. (2022). *Global status report on blood safety and availability 2021* (tech. rep.) (Licence: CC BY-NC-SA 3.0 IGO). World Health Organization. Geneva.

Zucoloto, M. L., Gonçalez, T. T., Gilchrist, P. T., Custer, B., McFarland, W., & Martinez, E. Z. (2019). Factors that contribute to blood donation behavior among primary healthcare users: A structural approach. *Transfusion and Apheresis Science*, *58*(5), 663–668. https://doi.org/https://doi.org/10.1016/j.transci.2019.08.020

## APPENDIX A

This section reports the Python libraries used in this study (Table 9) and further resources on implemented machine learning algorithms (Table 10).

Table 9: Python libraries utilized in the study

| Library | Version | Reference |
|---|---|---|
| Pandas | 2.2.2 | McKinney, 2010 |
| Numpy | 1.26.4 | Harris et al., 2020 |
| Seaborn | 0.13.2 | Waskom, 2021 |
| Matplotlip | 3.8.0 | Hunter, 2007 |
| Imblearn | 0.13.0 | Junder et al., 2018 |
| Scikit-learn | 1.5.2 | Pedregosa et al., 2011 |
| Xgboost | 1.7.6 | Chen and Guestrin, 2016 |
| Shap | 0.46.0 | Lundberg and Lee, 2017 |
| SciPy | 1.13.1 | Jones et al., 2007 |
| Missingno | 0.52.2 | Mikulski, 2023 |
| Pingouin | 1.17.0 | Vallat, 2018 |
| Statsmodels | 0.14.0 | Seabold and Perktold, 2010 |

Table 10: Machine learning algorithms implemented in the study

| Algorithm | Reference |
|---|---|
| XGBoost | Chen and Guestrin, 2016 |
| Random Forest | Breiman, 2001 |
| Multi-layer Perceptron | Rosenblatt, 1958, Rumelhart et al., 1986 |
| Support Vector Machine | Cortes and Vapnik, 1995 |
| Dummy Classifier | Pedregosa et al., 2011 |