

Computational intelligence methods for uncertain medical and sensor data

Ph.D. Dissertation

mgr inż. Olga Kamińska

Supervisor: dr hab. Katarzyna Kaczmarek-Majer

POLISH ACADEMY OF SCIENCES

Systems Research Institute

Warsaw, Poland

2025

Metody inteligencji obliczeniowej dla niepewnych danych medycznych oraz danych pochodzących z urządzeń sensorycznych

Rozprawa Doktorska

mgr inż. Olga Kamińska

Promotor: dr hab. Katarzyna Kaczmarek-Majer

Polska Akademia Nauk
Instytut Badań Systemowych
Warszawa, Polska
2025

Streszczenie w języku polskim

Wspomaganie lekarzy w opiece nad pacjentami dotkniętymi chorobami afektywnymi stanowi znaczące wyzwanie, gdyż tradycyjne metody modelowania danych stosowane w celu predykcji lub monitorowania stanu pacjenta nie uwzględniają wielu niepewności. Choroba afektywna dwubiegunowa to przewlekłe schorzenie psychiczne, charakteryzujące się ekstremalnymi wahaniami nastroju, obejmującymi epizody maniakalne i depresyjne. Ograniczenia istniejących systemów komputerowych skłoniły do poszukiwania bardziej solidnych technik monitorowania, przy czym metody inteligencji obliczeniowej okazały się obiecujące.

W tym kontekście niniejsza rozprawa doktorska podejmuje wyzwanie opracowania zaawansowane metody inteligencji obliczeniowej, wykorzystujące dane akustyczne do monitorowania zmian nastroju u pacjentów z chorobą afektywną dwubiegunową, ze szczególnym uwzględnieniem niepewności wynikającej z charakteru zebranych danych głosowych oraz danych pochodzących z ankiet psychiatrycznych.

Praca porusza zagadnienie selekcji istotnego zbioru cech akustycznych, które precyzyjnie charakteryzują określone stany afektywne, przy jednoczesnym zapewnieniu interpretowalności wyników oraz uwzględnieniu niepewności związanej z danymi i etykietami uzyskanymi od lekarzy. Aby sprostać tym wyzwaniom, rozprawa wprowadza metodę *“Uncertainty Aware Advanced Feature Selection”*, umożliwiającą selekcję cech istotnych dla oceny stanu pacjenta. Dzięki uwzględnieniu niepewności związanej z nadawaniem etykiet przez psychiatrów podczas etapu wyboru istotnych cech wzrasta dokładność klasyfikacji stanu pacjenta. Następnie zaproponowana metoda zatytułowana *“Explainable Compositional Multilayer Perceptron”* uwzględnia niepewność związaną z ekstrapolacją etykiety medycznej oraz wykorzystuje architekturę sieci neuronowej z dwoma warstwami wyjściowymi, odpowiadającymi za ocenę stanu pacjenta i jego symptomów. Integracja tej metody z technikami wyjaśnialnymi sprawia, że zaproponowane podejście jest nie tylko dokładne, ale także bardziej interpretowalne, co ma kluczowe znaczenie w zastosowaniach klinicznych.

Ponadto, w niniejszej rozprawie zaproponowano metodę podejmującą problem niekompletnego i niespójnego etykietowania poprzez ocenę podobieństwa między danymi oznakowanymi a nieoznakowanymi, co umożliwia bardziej efektywne wykorzystanie dużych i często niespójnych zbiorów danych. Opracowana metoda *“Semi-supervised Cluster Validation Agreement”* umożliwia weryfikację spójności skupień w sytuacji, gdy dostępne są dane częściowo oznakowane, biorąc pod uwagę odległość od danej etykiety. Ten krok ma kluczowe znaczenie, ponieważ wykazuje, że skupienia są istotne i mogą być interpretowane w szerszym kontekście badań.

Dissertation summary

Assisting psychiatrists in the care of patients affected by affective disorders poses a significant challenge, as traditional data modeling methods used to predict or monitor a patient’s condition do not account for numerous uncertainties. Bipolar disorder is a chronic psychiatric condition characterized by extreme mood swings, including manic and depressive episodes. The limitations of existing computer systems supporting this problem have led to the search for more robust monitoring approaches, with computational intelligence methods offering promising solutions.

In this context, the doctoral dissertation addresses the challenge of developing advanced computational intelligence methods using acoustic data to monitor mood changes in patients with bipolar disorder, with particular emphasis on incorporating the uncertainty resulting from the nature of the collected voice data and psychiatric survey data.

The proposed approach enables selecting a subset of acoustic features that accurately characterize specific affective states while ensuring the interpretability of the results and accounting for the uncertainty associated with the data and the medical labels. To meet these challenges, the dissertation introduces the “*Uncertainty Aware Advanced Feature Selection*” method, which is used for selecting features for assessing the patient’s condition. By taking into account the uncertainty associated with psychiatrists’ labeling during the feature selection stage, the accuracy of patient state classification is improved.

Next, the novel “*Explainable Compositional Multilayer Perceptron*” method presents a comprehensive approach to the problem, addressing the uncertainty related to the extrapolation of medical labels and employing a neural network architecture with two output layers—one for evaluating the patient’s condition and the other for assessing their symptoms. The integration of this method with explainable techniques makes the proposed approach not only comprehensively effective but also transparent, which is crucial in clinical applications.

Furthermore, this dissertation addresses the problem of incomplete and inconsistent labeling by evaluating the similarity between partially labeled datasets, enabling more efficient use of large and often inconsistent datasets. The proposed method “*Semi-supervised Cluster Validation Agreement*” verifies the consistency of clusters between labeled and unlabeled data by considering the distance from a given label. This step is crucial, as it demonstrates whether the clusters are relevant and can be interpreted in a broader research context.

Nomenclature

Abbreviation	Definition
AI	Artificial Intelligence
ANN	Artificial Neural Network
BD	Bipolar Disorder
CCFS	Cost-Constrained Feature Selection
CNN	Convolutional Neural Network
CVA	Cluster Validation Agreement
D	Depression (one of bipolar disorder affective states)
E	Euthymia (one of bipolar disorder affective states)
HMD	Hamilton Depression Rating Scale
KNN	K-Nearest Neighbor
LS	Linguistic Summarization
LSTM	Long Short-Term Memory
M	Mania (one of bipolar disorder affective states)
MLP	Multilayer Perceptron
NN	Neural Network
PLENARY	explaining black-box models in Natural Language through fuzzy linguistic summaries
RFE	Recursive Feature Elimination
ROC AUC	The area under the ROC curve
SCVA	Semi-supervised Cluster Validation Agreement
SHAP	SHapley Additive explanations
SL	Supervised Learning
SOM	Self-Organizing Maps
SSL	Semi-Supervised Learning
SVM	Support Vector Machine
UAAFS	Uncertainty Aware Advanced Feature Selection
UL	Unsupervised Learning
X	Mixed state (one of bipolar disorder affective states)
XAI	Explainable Artificial Intelligence
XCMLP(CF)	Explainable Compositional MLP (with Confidence Factor)
XGB	extreme Gradient Boosting
YMRS	Young Mania Rating Scale

List of Figures

1	Visualization of Bipolar Disorder	21
2	Example of bipolar disorder fluctuations over time	22
3	Example of Monarca application [7]	27
4	Example of Monsenso application	27
5	View of the main page in the application Moodmon.	29
6	Labels Confidence representation example	31
7	Possible strategies of labels uncertainties	32
8	Semi-supervised learning emplacement	39
9	Division of Feature Selection methods	44
10	Illustrated cost of obtaining features	50
11	Proposed basic ANN architecture for BD states classification	53
12	Proposed basic CNN architecture for BD states classification	54
13	Proposed basic LSTM architecture for BD states classification	55
14	A three-part approach called PLENARY [48] that represents compositional classification model, explanation of the predictive model's and linguistics summaries	57
15	The workflow of Explainable Compositional Multi-Layer Perceptron with Confidence factor (XCMLPCF)	58
16	Architectural diagram of the multi-task neural network in PLENARY [48].	59
17	An example of SHAP global explanation that present an impact of three acoustic features for hamd_anxiety classification	61
18	The workflow diagram presented in [51]	64
19	Degree of agreement for SOM for (left) with expert parameters and (right) automatic FS subset of acoustic parameters. [53]	65
20	Self-Organizing Map topologies considered during clustering and SCVA methods	67
21	Clustering with Self-Organizing Maps using Topology 1x3 preserving label uncertainty	68
22	SCVA heatmap	69
23	The data flow in <i>bipolar</i> package	73
24	Example of dashboards received by <i>bipolar</i> package.	74
25	Cost Constained feature selection results	82
26	Division of collected data into training, validation and test subsets	84
27	LOSS for 3 architecture	85
28	The AUC ROC for the CNN. Class 0 represents euthymia, 1 - depression, 2 - mania, 3 - the mixed state.	86
29	The loss function plot for the XCMLP model (green and red) and XCMLPCF model (blue and orange)	88

30	The ROC AUC curve for each of classes where 0 is euthymia, 1 is depression, 2 is mania and 3 is mix.	89
31	Confusion Matrix for those two considered methods.	89
32	The SHAP summary plot indicates an influence of the most important features for all of the classes.	92
33	The SHAP summary plot indicates an influence of the most important features for depression (left) and mania (right)	92
34	The SHAP summary plot indicates an influence of the most important features for all of the symptoms	93
35	The SHAP summary plot indicates an influence of the most important features for anxiety (left) and elevated mood (right)	93
36	The considered four types of SOM topologies used during clustering and SCVA method where patient received 2 equal medical assessment (Euthymia) during 2 visits.	95
37	The considered four types of SOM topologies used during clustering and SCVA method where patient received 2 different medical assessment (Mania and Depression) during 2 visits.	96
38	The heatmaps for considered pair of 2 visits for different types of topologies with result of SCVA value.	98
39	The heatmaps for considered pair of 2 visits for different types of topologies with result of SCVA value where parameter visit_ratio is above 0.5	98

List of Tables

1	Leading characteristics of Mania and Depression	23
2	Description of Bipolar Disorder Symptoms	24
3	Voice markers specification based on BD episodes and gender	25
4	Characteristic of data analysis and machine-learning approaches used for phase recognition and phase change detection	34
5	Frequencies of clusters presented in Figure 21	69
6	The list of scenarios that were created to verify the UAAFS method	76
7	List of important features for Depression with considered label con- fidence function	77
8	List of important features for Mania with considered label confi- dence function	77
9	Train and test sets data distribution	78
10	Classification results for 4 classes using XGB, SVM, and KNN	78
11	Classification results for 2 (euthymia vs depression) class using XGB, SVM, and KNN	79
12	Classification results for 2 classes (euthymia vs mania) using XGB, SVM, and KNN	80
13	Train, Valid and Test Data distribution	84
14	Detailed results for 3 different type of NN architecture. Details information like layers or hyperparameters are available in Section 5.1	85
15	Detailed results received with PLENARY architecture	87
16	The summary table with list of considered patients and combi- nation of visits using SCVA method. PID: patient ID, VComb: combination of patient visit, V1,V2: date of visit 1 and 2, L1 and L2: psychiatric assignment (label 0:Euthymia, 1:Depression, 2:Ma- nia, 3:Mixed), Obs1, Obs2: number of collected observation around Visit 1 and 2, Visit Ration: ratio between 2 number of collected observation around visits, SCVA: result for SCVA for 4 considered topologies.	99

List of author publications

1. **Kamińska O.** Kaczmarek-Majer K. and Opara K. and Jakuczun W. and Dominiak M. and Święcicki Ł. and Hryniewicz O. **Self organizing maps using acoustic features for prediction of state change in bipolar disorder** In: Marcos M. et al. (eds) Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems. KR4HC 2019, TEAAM 2019. Lecture Notes in Computer Science, vol 11979. Springer
2. **Kamińska O.** Kaczmarek-Majer K. and Hryniewicz O. **Acoustic features for prediction of state change in bipolar disorder.** Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2020, Lisbon
3. Kaczmarek-Majera K., Casalino G., Castellano G., Dominiak M., Hryniewicz O., **Kaminska O.**, Vessio G., Diaz-Rodriguez N., **PLENARY Explaining black-box models in natural language through fuzzy linguistic summaries** In: Information Sciences Volume 614, October 2022, Pages 374-399
4. **Kamińska O.** Kaczmarek-Majer K. and Hryniewicz O. **Impact of clustering unlabeled data on classification: case study in bipolar disorder.** Proceedings of the 17th Conference on Computer Science and Intelligence Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, D. Ślęzak (eds). ACSIS, Vol. 30. 2022, 931–934.
5. **Kamińska O.** , Klonecki, T., Kaczmarek-Majer, K. (2024). **Feature Selection in Bipolar Disorder Episode Classification Using Cost Constrained Methods.** In: Juarez, J.M., et al. Explainable Artificial Intelligence and Process Mining Applications for Healthcare. XAI-Healthcare PM4H 2023. Communications in Computer and Information Science, vol 2020. Springer, Cham.
6. Kaczmarek-Majer K., Dominiak M., Antosik A. Z. , Hryniewicz O., **Kamińska O.**, Opara K. and Radziszewska W., Sochacka M. , Święcicki Ł. **Acoustic features from speech as markers of depressive and manic symptoms in bipolar disorder: A prospective study** In: Acta Psychiatrica Scandinavica 2024
7. Kamińska D., **Kamińska O.**, Sochacka M., Sokół-Szawłowska M. , **The Role of Selected Speech Signal Characteristics in Discriminating Unipolar and Bipolar Disorders** In: Sensors 2024, 24(14), 4721;
8. Sokół-Szawłowska M, **Kamińska O.**, Sochacka M., **MoodMon: novel optimization of bipolar disorder monitoring through patient-driven voice parameter submission and AI technology** In: Advances in Psychiatry and Neurology/Postępy Psychiatrii i Neurologii (2024), 33.4, pp. 230-240;

Contents

Nomenclature	4
List of Figures	5
List of Tables	7
1 Introduction	12
1.1 Motivation	12
1.2 Scope and case study in monitoring sensor and medical data	13
1.3 Problem Statement	14
1.4 Objectives of the dissertation and thesis statement	16
1.5 Major Contributions	18
1.6 Structure of this dissertation	19
2 Medical background in mental health monitoring and related uncertainties	21
2.1 Introduction to Mental Health Monitoring	21
2.1.1 Mental Health Overview	21
2.1.2 Application for Mental Health Monitoring	25
2.2 Computational Intelligence challenges in Bipolar Disorder	30
2.3 Label confidence considered in Bipolar Disorder	30
3 Computational intelligence methods for medical sensor data	33
3.1 Review of Automatic Feature selection in Mental Health	33
3.2 Artificial Intelligence Methods for Medical Sensor Data	36
3.2.1 Classical Supervised Learning Methods	36
3.2.2 The Evolution of Neural Networks in Modern Supervised Learning	37
3.3 Semi-supervised methods for Medical Sensor Data respecting uncertainty of data	38
3.4 The role of explainability in mental health context	40
4 Feature selection	43
4.1 The Role of Feature Selection in Computational Intelligence	43
4.2 Uncertainty Aware Advanced Feature Selection	45
4.2.1 Method overview	45
4.2.2 Recursive Feature Elimination	47
4.2.3 Selected classification model: XGBoost	48
4.3 Cost Constrained Feature Selection	50

5	Explainable supervised learning in mental health monitoring	52
5.1	Classification of Bipolar Disorder states with selected Neural Networks architecture	53
5.2	Explainable Compositional MLP approach for the disease and its symptoms	57
6	Semi-supervised Clustering	62
6.1	The Role of Unsupervised Learning in Computational Intelligence .	63
6.2	The Cluster Validation Agreement	64
6.3	The Proposed Semi-supervised Validation Approach based on Cluster Agreement	67
7	Results	71
7.1	BDmon system and data transformation	71
7.2	Hardware and Software	72
7.3	Numerical results	75
7.3.1	Uncertainty Aware Advanced Feature Selection Method . . .	75
7.3.2	Cost-Constrained Feature Selection	82
7.3.3	Primary Neural Network Classification	84
7.3.4	Explainable Compositional Multilayer Perceptron	87
7.3.5	Semi-supervised Cluster Validation Agreement	94
7.4	Discussion	101
8	Conclusions	103
8.1	Future work	106
	References	108
	Appendix	116

Acknowledgment

I want to express my enormous gratitude to my supervisor, Dr Hab. Katarzyna Kaczmarek Majer, for her constant support and guidance during the research and all the work we did together, significantly contributing to my professional development. Additionally, I am grateful to Prof. Dr Hab. inż. Olgierd Hryniewicz, the head of the Department of Stochastic Methods, for creating an excellent research environment and directing my research in the right direction.

Great appreciation is due especially to Małgorzata Sochacka, who initiated projects on the prognosis of phase changes with patients suffering from bipolar disorder. Her commitment, knowledge, and experience are invaluable.

I also want to thank my colleagues from the BIPOLAR project, Kamil Kmita and Jakub Małecki. Working with you has been extremely important in developing my scientific achievements.

The psychiatrist from the Institute of Psychiatry and Neurology in Warsaw deserve sincere gratitude, for the opportunity to collaborate and support during the deep exploration of the nature of bipolar affective disorder.

Finally, I would like to express my gratitude to my family. Your continued support and indulgence is invaluable.

1 Introduction

1.1 Motivation

The inspiration for this dissertation comes from taking part in research projects aimed at developing computational methods for predicting the upcoming phase change of patients with bipolar disorder using data collected from smartphones and psychiatric assessments. The projects were called BDmon, Moodmon, and BIPOLAR and were carried out at the Systems Research Institute Polish Academy of Sciences in cooperation with partners, in particular, medical doctors from the Institute for Psychiatry and Neurology in Warsaw.

Bipolar disorder (BD) is a chronic mental health disease characterized by extreme mood swings, including manic and depressive episodes. Effective management of BD requires continuous healthcare attention and consistent patient monitoring to ensure early detection of mood changes and timely interventions. However, monitoring BD patients presents numerous challenges. Traditional methods, which often rely on subjective self-reports or sporadic clinical assessments, are neither robust nor fully reliable. This creates gaps in timely detecting and managing mood phase transitions, potentially delaying appropriate interventions.

Hopefully, Artificial intelligence (AI) methods are increasingly being explored as promising solutions to enhance the monitoring and prediction of mood changes in BD patients. The rapid development of AI techniques opens up new opportunities to analyze complex behavioral patterns and predict mood phase changes more accurately. One particularly valuable type of data in this context is acoustic data, which has shown great potential in reflecting patients' current mental state [43]. Subtle changes in speech patterns, tone, and rhythm can serve as non-invasive, continuous markers of mood shifts, offering a powerful tool for patient monitoring. For example, it was confirmed that depressed patients had shown lower prosody (accent and inflection) and lower pitch than healthy patients [16].

However, a key challenge remains to adequately address the inherent uncertainties that arise during the collection and interpretation of behavioral and physiological data. These uncertainties may stem from variations in

- patient behavior: we do not see the patient and do not know the circumstances of the phone calls
- environmental factors: we do not know under what circumstances the patient talked on the phone
- inconsistency of medical assessment: the patient's condition cannot always be determined with an appropriate time label in detail, and what is more,

the doctor may interpret the patient’s behavior in different ways, just as it may be difficult for the patient to determine his or her symptoms

Therefore, managing these uncertainties effectively is crucial at every stage of the monitoring and modeling process to ensure the reliability of the algorithms being developed.

The purpose of this dissertation is to support mental health monitoring problems by developing new computational intelligence methods. The developed novel methods aim at:

1. selection of a subset of most important acoustic features for multidimensional data integrating information from psychiatric surveys;
2. classification of bipolar disorder states focusing on both the accuracy of the approach and its explainability;
3. semi-supervised validation approach aiming to include the uncertainty included in the medical data and to assess similarity considering this uncertainty.

In this context, AI models can use labeled mood states when available while continuously learning from acoustic and other behavioral data in real-time. This approach enhances the model’s ability to predict mood phase transitions, even when new labels are not immediately available. The combination of AI and semi-supervised learning enables the system to improve its accuracy over time, becoming more adept at identifying subtle indicators of impending mood shifts in BD patients.

Moreover, the success of such AI-driven monitoring relies on close collaboration between healthcare providers and patients. By facilitating communication through AI tools, doctors can better understand patient states, and patients can become more engaged in managing their condition. This enhanced partnership between patients and clinicians, supported by real-time AI insights, ultimately leads to more effective healthcare and improved quality of life for BD patients.

1.2 Scope and case study in monitoring sensor and medical data

The motivations for this study are rooted in addressing real-world challenges encountered during the collection and analysis of sensor and medical data, particularly in the context of monitoring patients with bipolar disorder (BD). The foundation of this research is a clinical trial conducted in Warsaw in 2017-2019, involving BD patients who were provided with a mobile application designed to monitor

their acoustic data, specifically capturing and analyzing voice recordings to identify phase changes.

However, several technical issues arose during the trial highlighting the complexities of developing robust AI models in medical contexts. One major challenge was the incomplete and inconsistent data stemming from various external factors. Patients occasionally stopped using the app, either due to technical difficulties such as internet connectivity issues or lapses in app updates, or due to personal reasons. As a result, gaps in the continuity of data emerged, which can significantly impact the performance and reliability of AI models tasked with detecting mood changes based on acoustic signals.

Additionally, the sparsity of labeled data presented a challenge in this trial. Mood labels, which were critical for training the AI models, were few and far between, provided only during clinical assessments or through patient-reported outcomes. Moreover, these labels were often marked by uncertainties, stemming from variations in how clinical scales like the Hamilton Depression Rating Scale (HAMD) [38] and Young Mania Rating Scale (YMRS) [89] were interpreted and translated into mood states. For example, determining the validity range of a particular label (how long it should apply to the data surrounding it) is a complex and imprecise task, introducing noise and uncertainty into the training data.

Another technical problem is the high volume of voice recordings collected throughout the study. With a large dataset of acoustic recordings, it becomes difficult to determine which acoustic features are most relevant in predicting mood changes. Acoustic parameters such as pitch, tone, speed, and pauses all potentially carry meaningful information, but identifying which of these parameters are most indicative of mood shifts in bipolar disorder remains a significant challenge. The sheer volume of data also poses computational challenges, as not all recordings may be equally valuable or informative.

In the field of medicine, the effective implementation of neural networks is paramount; however, the interpretability of these models is equally critical. Explainable AI (XAI) facilitates the elucidation of how individual input features contribute to the model’s predictions. This level of transparency is essential for healthcare professionals, as it allows them to understand the influence of specific patient characteristics and clinical data on the outcomes generated by the model. Ensuring that the parameters and their interactions are comprehensible not only fosters trust in the deployed algorithms but also enhances clinical decision-making processes.

1.3 Problem Statement

Given challenges presented in the previous Section, in this dissertation we propose novel methods to address selected uncertainties that are occurring in the

available datasets and the mental health monitoring problem. I start with a brief presentation and formalization of the main problems addressed in this dissertation.

Let us consider a dataset of raw voice data collected for an individual patient. Each of patient's mobile recording is described by a matrix M_{rp} where each row represents an ordered (10ms) frame and each column represents a particular acoustic feature, such as pitch, energy, spectral entropy etc. An explanation of those features can be found in the Appendix.

$$M_{rp} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{l1} & x_{l2} & \cdots & x_{lk} \end{bmatrix} \quad (1)$$

$$r, p, l, k \in \mathbb{N}$$

$$x \in \mathbb{R}$$

where

- r is the selected mobile recording, and $r \in R$, where R is number of all available mobile recordings
- p is the selected patient with BD, and $p \in P$ where P is subset of all available patients;
- l is a predefined parameter about the length of the recording considered for analyses (ordered number of frame), and $l \in L$ where L is the maximum length of all available frames;
- k is the predefined number of acoustic feature, and $k < K$ where K is maximum number of all available acoustic features.

Processing of raw and extensive data could generate performance problems. In one of my research [49] various aggregation operations have been considered, including some basic ones such as mean, standard deviation, quartiles, etc.. In that work, the best aggregating methods turned out to be the mean and skewness. Let us now consider processing of aggregated voice data described by vector W .

$$W_{rp} = [a_1 \ a_2 \ \cdots \ a_n] \quad (2)$$

$$n \in \{1, 2, \dots, 516\} \subset \mathbb{N}$$

where

- a is aggregated values of mobile recording.

The advantage of this approach is that it allows to reduce the size of the dataset and focus on the most relevant acoustic features that are representative of bipolar disorder.

The primary challenge addressed is the classification of bipolar disorder states, a topic that is extensively documented in existing literature [20], [90], [2], [1], [64].

Besides acoustic features, the considered dataset includes outcomes of the psychiatric assessment described using 2 numerical rating scales: Hamilton Depression Scale (18 questions concerning depressive symptoms) [38] and Young Mania Scale (17 questions associated with the intensity of manic symptoms) [89]. Using adequate transformation [46], we consider 4-class classification problems corresponding to the four bipolar disorder states (mania, depression, euthymia, mixed).

It is worth mentioning, as a side note, that psychiatric labels could be modeled in various ways. Specifically, one can attempt to forecast the problem as a linear one, trying to determine the number of points on an appropriate scale. Next, we could treat it as a binary problem to classify stable and unstable states. Alternatively, a multi-modal classification approach can be taken, where a specific class is identified. However, this disease has a multidimensional nature, so it can also be treated as a multi-class problem. In this dissertation, we will primarily focus on determining BD states, although in other works where I have had the opportunity to be a co-author, different approaches have been considered, see, e.g., [48, 52, 79]

Particularly, the main problem addressed in this dissertation is the selection of $k < K$ important acoustic features characterized in particular BD state aiming at both the accuracy of classification and its explainability. The secondary problem addressed in this dissertation is assessment of the similarity between partially labeled sensor data. This problem is mainly related to the lack and uncertainty derived from medical (psychiatric) labeling.

1.4 Objectives of the dissertation and thesis statement

The main research objectives of this dissertation focus on addressing the main and secondary problems presented in the previous Section through the implementation of advanced methodologies tailored to the specific nature of acoustic data in medical applications.

The research objective of this dissertation is to effectively address the selection of a critical subset of k acoustic features that accurately characterize specific bipolar disorder (BD) states taking into account the imprecision of the psychiatric labeling process, with an emphasis

on interpretability. Selecting appropriate features is crucial for constructing a robust and reliable predictive model, especially in the medical field, where data collection is often constrained by practical limitations. In medical contexts, feature selection is particularly challenging and essential due to the frequent limitations on variable availability. The inherent uncertainty of acoustic data and their labels must be carefully considered during the feature selection process. Interpretability is crucial in medical applications, where understanding the influence of each selected variable on outcomes, such as classification, is of utmost importance. To solve that problem the UAAFS method (described in Section 4.2) will be applied. That method, which accounts for label uncertainty during the selection of the k most important features, utilizing aggregated data that characterize the collected recordings. Furthermore, the problem of feature selection and feature explanation will be modeled using the XCLMPC method (described in Section 5.2). This method leverages a neural network to process raw acoustic data and, through the application of the SHAP model, provides an explanation of the selected features.

The secondary research objective of this dissertation is to effectively determine the similarity between partially labeled datasets to verify cluster coherence for the same affective states of bipolar disorder patients. The challenge of managing large acoustic datasets is exacerbated by the uncertainty in the data, which may result from uncontrolled collection conditions due to external factors, such as data transfer issues, memory limitations, loss of Internet connectivity, or unintended interruptions in smartphone use. Another significant challenge in real-world datasets is the lack of continuity of labels. While labels can be extrapolated to some extent, doing so introduces considerable uncertainty. To address that problem the method SCVA (described in Chapter 6) will identify the similarity between clusters formed around the same visits. This method aims to make effective use of unlabeled data by assessing their similarity to labeled data, even in the presence of uncertainty.

We pose the following three theses:

1. **The proposed Uncertainty Aware Advanced Feature Selection method effectively addresses the uncertainty arising from psychiatric label assignment over time, ensuring the selection of relevant and adequate acoustic features for the improved classification of bipolar disorder states.**
2. **The proposed Explainable Compositional Multi-Layer Perceptron with Confidence Factor (XCMLPCF) method enables us to adequately reflect the uncertainty related to affective symptoms. The proposed two-layer and multi-output approach enables both, the prediction, but also explanation of features that have an impact for**

the resulting state and symptoms, that is particularly important in the medical field.

3. **The proposed Semi-supervised Cluster Validation Agreement (SCVA)** enables to effectively determine the degree of similarity of unlabeled data to labeled data considering the temporal nature of sensor and psychiatric data collection. Thus, it is a promising approach to support the model validation process.

1.5 Major Contributions

Major contributions for that dissertation are following:

1. **Uncertainty Aware Advanced Feature Selection (UAAFS)** - is a method including label uncertainty into automatic feature selection in the context of acoustic data. In Section 4, there is a particular emphasis on the wrapper feature selection method based on acoustic data, aiming to identify which acoustic parameters are most significant in tracking mood changes. That approach is improved to focus on uncertainty arising from label assessment by using the XGBoost model. By refining the selection of acoustic features and enhancing the model's ability to focus on the most informative attributes, the accuracy of predictions can be further improved. The proposed UAAFS method is widely described in Section 4.2.
2. **Explainable Compositional Multi-Layer Perceptron with Confidence Factor (XCMLPCF)** - classification method for bipolar state with its explanation of acoustic features. The proposed XCMLPCF framework not only predicts the patient's condition but also indicates what factors influenced it and takes into account its uncertainty. That method is widely described in Section 5.2.
3. **Semi-supervised Cluster Validation Agreement (SCVA)** - validation method that inspects the consistency of clusters at selected time intervals while maintaining information about the certainty of labels. SCVA enables us to verify that patients behave congruously in similar groups. It validates the data (that were clustered firstly) that are similar in particular BD states. It could be a first step to create a monitoring system for BD patients aiming verify that some deviations in the data are related to a particular BD state. That method is widely described in Section 6.
4. Experimental evaluation for real-life datasets - during the study, the code repository was created. The source code includes exploratory data analysis connected with data preprocessing, and the next mentioned innovatory

methods are implemented. Repositories are available under the link placed in Section 7.2.

5. Literature review of monitoring application for mental health disorder (Section: 2.1.2), Computational Intelligence challenges and methods in bipolar disorder (Section: 2.2) and for medical and sensor data (Section: 3.1))

Besides the mentioned contribution, this dissertation covers the following aspects:

- importance of monitoring of patients in bipolar disorder;
- handling with acoustic data in mental health;
- approach to model uncertainty associated with assessing the labels;
- cost-constrained feature selection that cost of assessing features is raised;
- an overview of related literature on various computational and machine learning approaches (including neural network architectures, that support the classification of BD states and symptoms.

1.6 Structure of this dissertation

This dissertation has been arranged in such a way that it is a cross-section of the important steps of data analysis to support mental health monitoring, where the central point is to consider the uncertainty resulting from the specificity of the study and the data.

A brief introduction to each section is as follows:

Section 1: *Introduction* This chapter introduces the motivation, problem statement, objectives, and scope of the research.

Section 2: *Medical Background in Mental Health Monitoring and Related Uncertainties* An overview of mental health monitoring, with a particular focus on Bipolar Disorder, is provided. The chapter discusses the challenges associated with computational intelligence in mental health, processing of sensor and psychiatric data, and uncertainties in labeling Bipolar Disorder data.

Section 3: *Computational Intelligence Methods/Challenges for Medical Sensor Data* This chapter explores the application of computational intelligence in processing medical sensor data, particularly in mental health. It discusses automatic feature selection, neural networks, semi-supervised methods, and the role of explainability when working with mental health data.

Section 4: *Feature Selection* Feature selection methods relevant to mental health data are discussed, including uncertainty-aware advanced feature selection (UAAFS).

Section 5: *Semi-supervised Clustering* The role of unsupervised learning in computational intelligence is reviewed, with a specific focus on cluster validation. The chapter presents a SCVA methods.

Section 6: *Explainable supervised learning in mental health monitoring* Selected methods for classification in computational intelligence are reviewed, followed by the presentation of the Explainable Compositional MLP (XCMLPCF) approach for understanding disease and its symptoms.

Section 7: *Results* This chapter provides a comprehensive overview of the experimental results, including the design of experiments, datasets used, and numerical results. The performance of the proposed methods (UAAFS, CCFS, XCMLPCF, SCVA) is evaluated.

Section 8: *Conclusions* The final chapter summarizes the key findings of the dissertation, reflects on the contributions made, and suggests future research directions.

At the end, a list of references cited throughout the dissertation and, additionally, there is an Appendix with a list of available acoustic features along with their explanation.

2 Medical background in mental health monitoring and related uncertainties

2.1 Introduction to Mental Health Monitoring

2.1.1 Mental Health Overview

Mental health is a crucial aspect of overall well-being, influencing how we think, feel, behave, and manage stress. Mental health disorders, which affect a person's thinking, mood, and behavior, can significantly impact one's life. These disorders result from a complex mix of genetic, biological, environmental, and psychological factors. Common examples include Depression, PTSD, Schizophrenia, Anxiety Disorders, and Bipolar Disorder.

Mental health is one of the significant challenges to societies and healthcare budgets. According to the World Health Organisation (WHO) mental illness will become the largest health burden for society in the years to come.

- Every year, 38% of the population in Europe is affected by a mental disorder ([86]);
- According to WHO estimated 40 million people live with bipolar disorder worldwide.

Subgroup of Mental Health Illnesses are Affective disorders, also known as mood disorders, which are a category of mental health conditions characterized by significant disturbances in a person's emotional state. In ICD10 (International Classification of Diseases) list by the WHO it is placed in F30-F39. These disorders primarily impact mood, which can fluctuate between extremes of low mood (depression) and high mood (mania or hypomania). They can interfere with daily functioning, relationships, and overall well-being if left untreated.

The two most common types of affective disorders are major depressive disorder (MDD) and bipolar disorder (BD). Major depressive disorder is marked by prolonged periods of sadness, feelings of worthlessness or guilt, fatigue, loss of interest in activities, and sometimes thoughts of suicide. It is one of the most prevalent mental health conditions worldwide and can range from mild to severe.

Bipolar disorder (BD) [65] on the other hand involves mood swings that alternate between depressive episodes and manic or hypomanic episodes. During a manic phase, individuals may experience heightened energy, euphoria, racing



Figure 1: Visualization of Bipolar Disorder

thoughts, decreased need for sleep, and impulsive or risky behavior. In a depressive phase, symptoms are similar to those of MDD, including deep sadness, hopelessness, and fatigue. Bipolar disorder is divided into Bipolar I, which includes full manic episodes, and Bipolar II, which involves milder hypomanic episodes. These fluctuations can affect sleep, energy, behavior, judgment, and clarity of thought [81]. Bipolar disorder is lifelong and requires ongoing care to manage symptoms effectively. Diagnosis of bipolar disorder is challenging due to overlapping symptoms with other conditions like depression and anxiety. Accurate diagnosis is critical for effective treatment, which may involve mood stabilizers, psychotherapy, and careful monitoring to prevent triggering manic episodes through inappropriate medication, such as antidepressants. In Figure 2 is shown an example of available mood fluctuation over time.

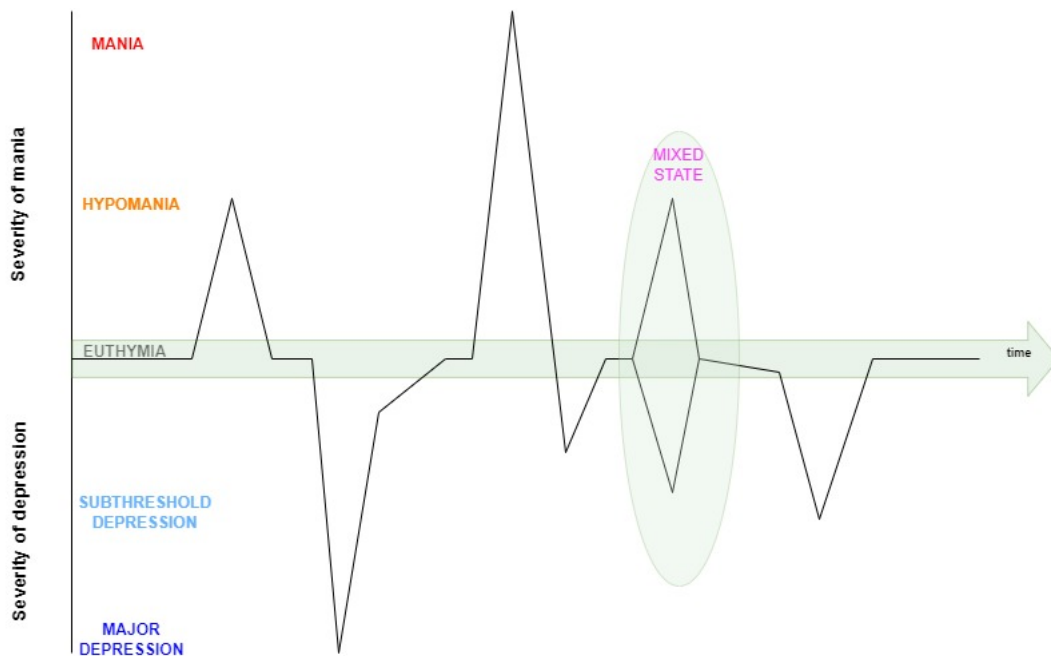


Figure 2: Example of bipolar disorder fluctuations over time

Characteristics of Bipolar Disorder States The disorder involves episodes of euthymia, mania, and depression (both presented in Table 1) and mixed state, where each episode presents distinct symptoms.

Table 1: Leading characteristics of Mania and Depression

Mania Symptoms	Depression Symptoms
Elevated mood	Persistent sadness
Increased energy	Fatigue
Racing thoughts	Loss of interest in activities
Decreased need for sleep	Difficulty concentrating
Impulsive behavior	Feelings of worthlessness or guilt
Euphoria	Changes in appetite or weight
Irritability	Suicidal thoughts
Grandiosity	Hopelessness

Mixed Episodes is a dangerous combination of manic and depressive symptoms that could occur simultaneously. Feeling sad yet highly energized can increase the risk of suicide.

Euthymia in Bipolar Disorder is a state of stable mood, where individuals experience neither mania nor depression. During euthymic periods, those with bipolar disorder maintain a balanced mood, normal energy levels, improved cognitive functioning, stable relationships, and a higher quality of life. Achieving and maintaining euthymia is the primary goal of treatment, which involves mood stabilizers, therapy, and lifestyle management.

Symptoms The class labels corresponding to mental states were assessed using 18 questions related to depressive symptoms from the Hamilton Depression Rating Scale (HAMD) and 17 questions concerning the intensity of manic symptoms from the Young Mania Rating Scale (YMRS).

However, questions from both scales could be grouped into specific **symptoms**, as detailed in Table 2, and these symptoms could be treated as intermediate labels in BD. The first column of the table describes the symptom group, the middle column provides a detailed description of each symptom group along with the related questions to be evaluated, and the last column indicates the maximum number of data points for each scale. The more data points assessed, the more intense the symptoms are considered within a given group for a specific patient.

Introducing symptoms as an intermediate level of annotations for bipolar disorder (BD) data is a key innovation of this work, as it has not been previously applied

in the context of AI decision support systems for monitoring mental disorders. Additionally, it has been observed that low-level acoustic data can be challenging for psychiatrists to interpret effectively.

Table 2: Description of Bipolar Disorder Symptoms

Symptom	Description	Max points
Anxiety	Decrease anxiety, fear mental symptoms	6
Decreased activity	Inhibition, work, and interests	8
Decreased mood	Criticism, depressing mood, feeling guilty	10
Disorganization	Appearance, formal thoughts disorder, thoughts disorder, view	16
Elevated activity	Increased activity, speech	12
Elevated mood	Elevated mood	4
Irritability	Irritability, destructive behavior	16
Sleep disorder	Early awakening, intermittent sleep, sleep disorder, sleep	10
Somatization	Fear somatic symptoms, generic somatic symptoms, hypochondria	10
Suicide	Suicide tendencies	4

Recent research [47] discriminating the relation of acoustics features with particular BD symptoms. Loudness, pitch, voice signal spectrum, voice quality, and speaking time significantly correlate with the severity of depressive and manic symptoms. What is more authors confirm that acoustic parameters extracted from daily phone calls are useful for predicting of bipolar disorder phase.

Voice markers discriminating BD states In 2024 a paper [47] appeared when authors investigated the relation between the mentioned symptoms in BD. They grouped a set of acoustic features into 3 categories:

1. **Prosodic features** that reflect changes over longer segments of time, perceived as rhythm and melody of speech, mainly expressed through pitch, energy, loudness, speaking rate, and pauses
2. **Spectral features** reflecting the spectrum of speech and the frequency of the distribution of the speech signal at a given time, for example, melcepstral coefficients, fundamental frequency and its harmonics, and so on
3. **Voice quality-related features** reflecting airflow and resonant properties through the glottis, vocal fold vibrations, for example, jitter, shimmer, harmonics to noise ratio, and so on.

According to that paper, we could distinguish the main conclusion in the scope of BD symptoms in Table 3:

Table 3: Voice markers specification based on BD episodes and gender

	Gender	MANIAC symptoms	DEPRESSION symptoms
Prosodic Features	male	higher pitch, louder and more energetic speech	lower tone of voice, less energetic speech
	female	clarity of speech decreased with increased mania	No clear correlations
Spectral Features	male	clearer speech, sharper sound	less clearly speech , slurred speech
	female	less clearly	no correlation
Quality Features	male	correlated with rougher and more irregular voices	less rough and irregular voice
	female	less rough and more regular voices	rougher and more irregular voices

Based on Table 3 we can conclude that mostly the characteristics of women and men are opposite and there are differences between maniac and depression episodes.

2.1.2 Application for Mental Health Monitoring

The Mental Health Monitoring App is a comprehensive tool designed to help users track, manage, and improve their mental well-being. This application caters to a broad audience, from those experiencing mild stress to individuals managing more severe mental health conditions. It offers a range of features that support users in understanding their mental state, providing tools for self-care, and offering professional support options.

In fact, there is an exponential increase (more than 10 000 [84]) in the number of mental health apps available on the internet, with less than 5% being actually studied. This meta-review revealed that apps for anxiety and depression hold great promise with clear clinical advantages, either as stand-alone self-management or as adjunctive treatments. More meta-analyses and more quality studies are needed to recommend apps for other mental health issues or for specific populations.[58]

The development of mental health monitoring applications has been shaped by technological advancements, increasing mental health awareness, and a growing demand for accessible care.

The field of digital health has seen remarkable advancements over the past decade, particularly in the development of tools and platforms designed to help individuals manage bipolar disorder. From mood tracking apps to sophisticated AI-driven platforms, these initiatives have significantly improved the way bipolar disorder is monitored and treated.

Smartphone-based interventions and monitoring have a significant positive impact on depressive and manic symptoms of BD patients in between-group and within-group analysis [62]

MONARCA (2009 - 2013) [7] was one of the pioneering projects that used mobile technology to help individuals with bipolar disorder manage their condition. Developed through a European research initiative, MONARCA provided tools (see Figure 3)for mood tracking, medication adherence, and sleep monitoring, with data sharing capabilities that allowed healthcare providers to offer more personalized and timely interventions. MONARCA integrates self-assessment via a smartphone app and a web portal for patients and clinicians. It collects sensor data, tracks medication, and provides a historical overview for personalized care. This project laid the groundwork for the next generation of digital mental health tools. The MONARCA system helps by supporting self-management through features like self-assessments, activity monitoring, coaching, and data sharing. Users found the system easy to use, appreciating its convenience compared to traditional methods.

Building on the success of MONARCA, **Monsenso** [66](2016 - Present) emerged as a commercial platform that significantly expanded MONARCA’s capabilities. It is a clinically validated digital health solution designed for managing mental health disorders. Monsenso integrates data from mobile apps, wearables, and other health devices, offering real-time monitoring, personalized feedback, and seamless communication with healthcare providers. This platform has become a cornerstone in the clinical management of bipolar disorder, bridging the gap between patients and their healthcare teams.

Monsenso is a clinically validated digital health solution designed for managing mental health disorders. It provides tools for patients, clinicians, and caregivers to monitor symptoms, track progress, and optimize mental health treatment. The platform aims to improve mental health accessibility and outcomes by integrating self-assessment data, clinician tools, and personalized content delivery.

The Monsenso platform collects and aggregates data from patient self-assessments and passive sensor data. The system offers predictive analytics, mood forecasting, and correlation analyses between behaviors and illness progression. This data is used to assess treatment effectiveness and anticipate patient needs. However, advanced statistical tools like pattern analysis and risk assessments are limited to research purposes. Monsenso is certified as a Class 1 CE medical device and adheres



Figure 3: Example of Monarca application [7]

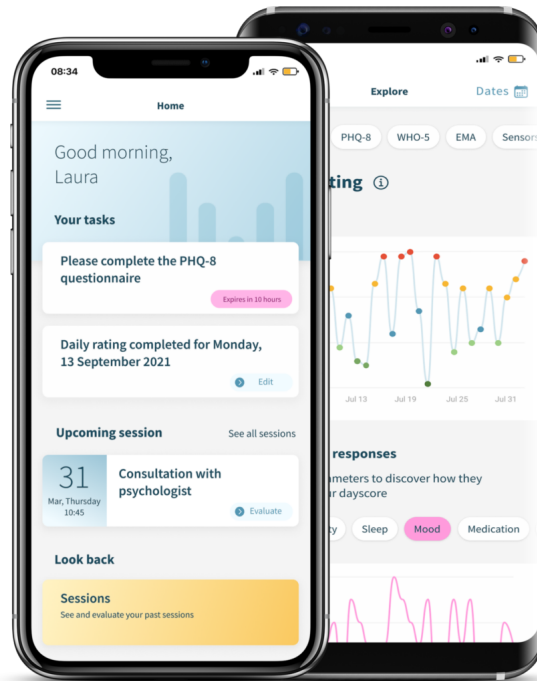


Figure 4: Example of Monsenso application

to HIPAA regulations for data security.

Besides MONARCA and Monsenso, many other systems were helpful with managing mental health problems. Some of them deserve attention.

RADAR-CNS (2016 - Present) [73] is a large-scale European project that uses wearable devices and mobile technology to monitor and prevent relapses in bipolar disorder, depression, and epilepsy. By collecting real-time data and applying machine learning algorithms, RADAR-CNS aims to predict relapses before they occur, offering a proactive approach to managing mental health conditions. That project was one of the largest remote disease-monitoring studies in Europe [17] that collected a unique multilingual clinical speech dataset.

BEHAPP (2017 - Present) [80] is a research platform that focuses on passive data collection through smartphones to support scientific and medical research. Operated by the Behapp research team at the University of Groningen (Netherlands), this service uses a smartphone app to gather data without recording any personal content (like messages or conversations). Depending on the study, the app may collect various types of data such as location, call and SMS history (with anonymized numbers), WiFi access points, app usage, motion through accelerometers, ambient light, and noise levels. Device information is also recorded for quality control, but no unique device identifiers are stored.

MHDeep (2018 - Present) [39] is a framework that leverages deep learning techniques to analyze complex mental health data, with a focus on improving the diagnosis of various mental health disorders and management of mood disorders, including bipolar disorder. The system transforms data obtained from sensors integrated in a smartwatch and smartphone. Results received by the framework are around even 100% to recognize between mental disease and healthy people using about 30 minutes of recording. MHDeep is computationally efficient, making it suitable for continuous diagnosis and daily monitoring.

BiAffect (2017 - Present) [91] introduced a novel approach to mental health monitoring by analyzing typing dynamics on smartphones to detect mood changes in individuals with bipolar disorder. By monitoring factors like typing speed, rhythm, and errors, BiAffect provides an innovative, passive method for early detection of mood shifts, allowing for timely intervention.

Finally, the project “Smartphone-based diagnostics of phase changes in the course of bipolar disorder” (**BDmon**) - conducted in 2017-2019, among others, by the Systems Research Institute PAN and by the Institute of Psychiatry and Neurology in Warsaw. The subjects of the analysis are processed acoustic data recorded from patients from the bipolar study to predict the upcoming phase change in a patient with bipolar disorder. Data were collected during daily telephone conversations what makes this subset unique and valuable. Detailed information about that project is presented in Section 7. BDmon as mentioned before Monarca has

obtained a successor called **Moodmon**, firstly described in [52]. The scope of that clinical study project is to forecast forthcoming changes in Bipolar Disorder using telephone surveys that examine the speech signal on a given day and wristbands. Besides voice analysis, patients with Bipolar Disorder and Unipolar Disorder (UD) are observed by the psychiatrist at least once a month mostly by telephone visits to assess their current mood. Visualization of mobile application is presented in Figure 5. The first forecasting results in that project are described in [79].

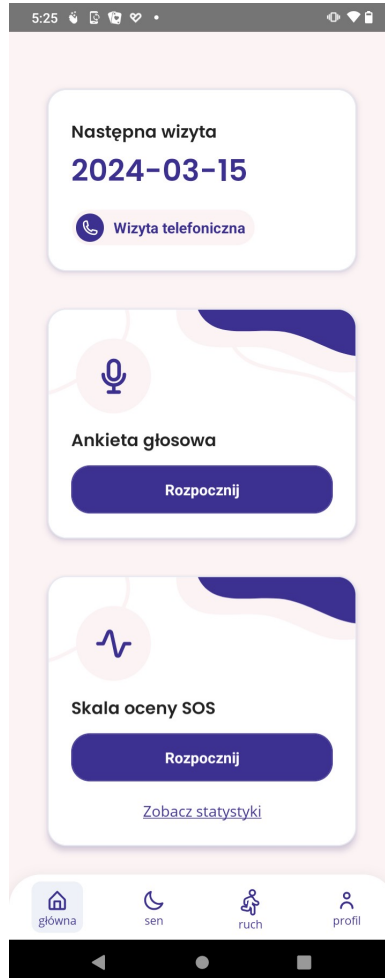


Figure 5: View of the main page in the application Moodmon.

2.2 Computational Intelligence challenges in Bipolar Disorder

By reviewing the latest research on bipolar disorder, we can distinguish 2 areas that are widely described and developed:

Diagnosis Diagnosing bipolar disorder is often delayed and may not be confirmed until after many years of living with the condition [8]. In the current trend of diagnostic imaging, techniques like MRI (Magnetic Resonance Imaging) and EEG (Electroencephalography) are frequently used to study and detect anomalies associated with bipolar disorder. In the literature, several studies [88] [21] [60] rely on EEG signals to differentiate bipolar disorder patients from healthy individuals, while others use MRI imaging [70] [61] [78] to examine brain characteristics associated with that disorder. Authors mainly used Convolutional Neural Networks [55] for that studies as the first choice dealing with images.

Predictive Modeling for Mood Episodes and Real-time Monitoring One of the most precious is predictive modeling for mood episodes that is that is combined with Real-time monitoring. It refers to using computational techniques, to forecast the likelihood of mood swings or episodes (such as manic or depressive episodes). The goal is to create mostly classification models based on time series data coming from different sensors that can accurately predict when a mood episode might occur, allowing for timely interventions that can help prevent or mitigate the severity of the episode. It is essential to collect data in real-time to be able to track changing patient data depending on the day and react as soon as possible if necessary.

The best example for that purpose is mentioned in Section 2.1.2 application named Monarca [7] and its successor Monsenso [66]. Moreover in Sections 2.1.2 and 7 there are described widely applications BDmon and Moodmon that were prepared for resolving that problem.

Both problems of feature selection and determine similarity between data raised in that dissertation have a significant impact on that scope.

2.3 Label confidence considered in Bipolar Disorder

The process of assigning labels is accompanied by several uncertainties most of all annotating the audio recordings with class labels. One issue is that the method of interpreting scores from the Hamilton and Young questionnaires has strictly defined value ranges, so sometimes a difference of just one point can completely change the classification of a patient's state. Label Confidence Factor is a metric that uses transformed labels (just like psychiatric assessments) described by L_{pt} .

The time valid for each assessment is valid for the current day of observation t_0 . That assumption entails that total number of labeled data could be negligible. However in related works [68], [46] the ground truth period refers to the fixed time window surrounding the visit which is from 7 days before the visit (t_{-7}) up to 2 days after the visit (t_{+2}). It means that the number of labeled data increases significantly as presented in Figure 6. The dotted line indicates the phone calls of the patient, filled rectangle indicates the day of assessment, and the blank rectangle indicates the period that we are extrapolating. As a result, all data within this period are categorized uniformly.

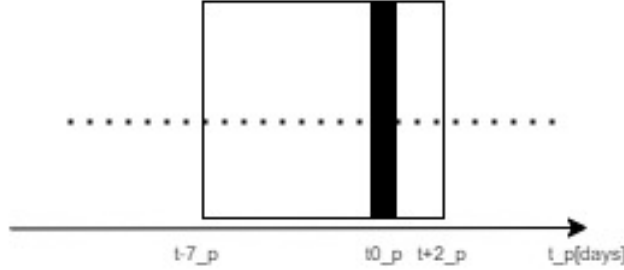


Figure 6: Labels Confidence representation example

Moreover, despite the assumption that most phone calls within the ground truth period exhibit specific characteristics of the disease phase, we lack direct labels specifically assigned to these supervised calls. The ground-truth based method has the potential to lead to misleading oversight, which can impact the performance of feature selection algorithms and later, the performance of the predictive model. Therefore, we address the issue of label uncertainty by assuming that the labels are precise but may be assigned to individual observations with varying levels of confidence.

Taking into account the varying confidence about labeling already at the feature selection stage is challenging and have a higher impact to considered data set. The confidence factor could be simulated as various scenarios and leads to solving the problem of selected important features in current dissertation.

The goal is to extrapolate the state from a visit date to a broader time window and introduce variability. We'd also like to quantify the confidence of our extrapolation. We will model how confident we are when extrapolating the actual patient's state on the time range we defined. For example, we may assume that we are most confident about the state in day 0, i.e. the visit day. Moving away from the visit day our confidence may decrease. This is one of the possible strategies depicted in Figure 7, where orange color corresponds to *constant* strategy - the state-of-the-art solution. Green color represents the *step* strategy, and blue color

represents the *custom* strategy in a form of a normalized Gauss curve parametrized by $\mu = 0, \sigma = 1$.

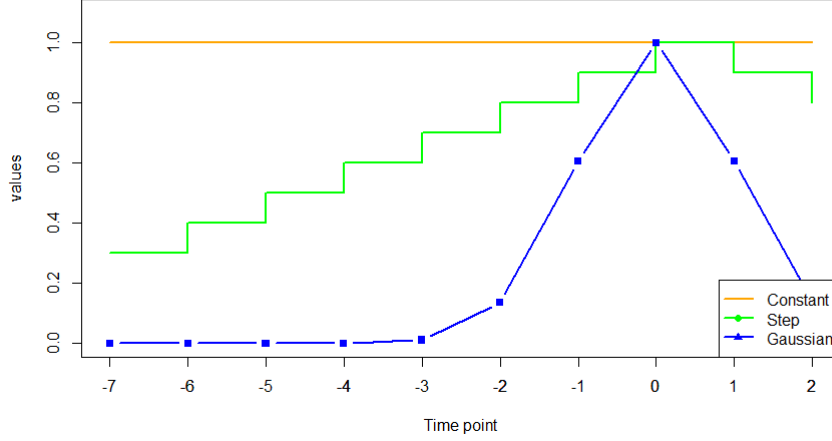


Figure 7: Possible strategies of labels uncertainties

The proposed labels confidence factor is based on [56], where the step functions were used to introduce varying levels of confidence assigned to labels obtained from the indirect annotation process.

The important aspect is that the confidence factor is always expressed as a number between 0 and 1. In the proposed approach, situations may arise where labels overlap because the patient meets with the psychiatrist at intervals shorter than a week. In such cases, we replace the collected values based on the higher Confidence Factor.

Clearly, both strategies the *gauss* and *step* weigh the observations recorded farther away from the visit with lower confidence. For example, any observation recorded on day -5 would get $\text{conf}_j^{\text{step}} = 0.5$ and $\text{conf}_j^{\text{custom}} = 0.00266$.

Modeling the uncertainty about the psychiatric labeling from the initial steps of the data processing and analysis is the key objective of this dissertation. Label confidence will be integrated and further investigated in Sections 4, 5 and 6.

3 Computational intelligence methods for medical sensor data

3.1 Review of Automatic Feature selection in Mental Health

The selection of important features in healthcare is often done using the expert method [28, 71, 67].

In 2020, in the International Journal of Medical Informatics, the review article named 'Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms, and predictive modeling' [5] was published. The idea of that article is to overview and discuss the studies on the smartphone-based systems that monitor or detect the phase change in BD. This article does not address the issue of feature selection; however, based on its content, I would like to review the included works once again to verify whether the feature selection method is automatic or expert-driven.

Published studies were identified by searching MEDLINE, PsychINFO, and Embase in March 2019. Follow-up citations were checked for additional relevant studies. Machine-learning-related studies were searched in Scopus in May 2019 using terms like "bipolar" or "BD" combined with "smartphone," "mobile," or "phone," limited to papers between 2000 and 2019. Articles were reviewed per PRISMA guidelines. Two raters screened de-duplicated abstracts, excluding non-English papers. Only mHealth studies were included. After full-text review of 37 papers, systematic reviews and study protocols were excluded. Discrepancies were resolved by a third rater.

Based on that article that Anna Z. Antosik-Wójcińska proposed in [5] I extended Table 4 with papers between 2020-2024 and complete with missing criterion named 'Type of feature selection' that will focus on approach to feature selection. The newest results are presented in Table 4.

Table 4: Characteristic of data analysis and machine-learning approaches used for phase recognition and phase change detection

No	Author	Year	Self-Monitored Data	Objective Data	Voice Data	Problem Formulation	Feature Selection
1	Alvarez-Lozano et al. [4]	2014	Yes	Yes	No	Phase change detection	Spearman's rank correlation coefficient.
2	Beiwinkel et al. [9]	2016	Yes	Yes	Yes	Identification of clinical relevance of smartphone data	using all of collected data
3	Bonsall et al. [10]	2012	Yes	No	No	Modeling interepisodic mood instability	analysed the full dataset
4	Bopp et al. [11]	2010	Yes	No	No	Identification of clinical relevance of smartphone data	using all of collected data
5	Cao et al. [12]	2017	Yes	No	No	Phase detection	using all of collected data
6	Depp et al. [15]	2012	Yes	No	No	Identification of validity of paper-and-pencil mood charts	using all of collected data
7	Depp et al. [18]	2015	Yes	Yes	No	Identification of validity of paper-and-pencil daily mood charts	using all of collected data
8	Faurholt-Jepsen et al. [28, 24]	2016	Yes	Yes	Yes	Phase detection	using all of collected data
9	Gideon et al. [30]	2016	No	No	Yes	Phase detection	using all of collected data
10	Grunerbl et al. [34]	2015	Yes	Yes	Yes	Phase change detection	filter feature selection method.
11	Guidi et al. [35]	2017	No	No	Yes	Phase detection	authors selection
12	Hidalgo-Mazzei et al. [41]	2017	Yes	Yes	No	Evaluate the potential of smartphone app to regulate biological rhythms	don't indicate how they select features
13	Karam et al. [54]	2014	No	No	Yes	Phase detection	greedy forward feature selection
14	Maxhuni et al. [64]	2016	Yes	Yes	Yes	Phase detection	empirical/combinational expert method, study different datasets with different groups of features

15	Muaremi et al. [68]	2014	Yes	No	Yes	Phase detection	filter feature selection method based on joint mutual information (JMI)
16	Osmani et al. [71]	2013	Yes	Yes	No	Phase change detection	using all of collected data
17	Vazquez-Montes et al. [85]	2018	Yes	No	No	Mood stability	using all of collected data
18	Moraga et al. [67]	2024	Yes	Yes	No	explore mood states' dynamics in patients with BD.	expert knowledge
19	Faurholt-Jepsen et al. [26]	2024	Yes	Yes	Yes	Patients classification (BD versus unipolar disorder)	using all of collected data

Based on Table 4, we can conclude that the overwhelming majority of the referenced studies either used the entire dataset or selected key features based on expert knowledge. Filter methods, using factors such as Spearman’s correlation coefficient or the Joint Mutual Information Coefficient, were employed in several studies. The wrapper method, however, was only applied once, as reported in [54].

3.2 Artificial Intelligence Methods for Medical Sensor Data

3.2.1 Classical Supervised Learning Methods

Classical methods in supervised learning are built on robust mathematical and statistical frameworks, offering interpretable and effective models in diverse applications. These models often require explicit feature engineering, a process that leverages domain expertise to transform raw data into meaningful representations. Despite the emergence of more complex models, classical techniques remain indispensable due to their computational efficiency and theoretical grounding. We could distinguish several methods of classical supervised learning focus on classification that are extensively described in [40]:

Logistic Regression, is utilized for binary classification tasks. It models the probability of an observation belonging to a particular class by applying the logistic sigmoid function to a linear combination of the input features. Despite its linear nature, logistic regression can be extended to multi-class classification via techniques like one-vs-rest or softmax regression.

Decision Trees are non-parametric models that recursively partition the input space into sub-regions based on feature values. At each split, the algorithm selects the feature and threshold that maximize a given criterion, such as information gain or Gini impurity. Decision trees are intuitive and interpretable, though they are prone to overfitting when applied to complex datasets.

Ensemble methods such as Gradient Boosting Machines (GBMs) and XGBoost build on the principles of decision trees. The extended description is available in Section 4.2.3.

Support Vector Machines (SVMs) are powerful supervised learning models used for both classification and regression tasks. SVMs work by identifying a hyperplane that maximally separates data points of different classes in the feature space. The goal is to maximize the margin between the nearest data points (support vectors) of each class, thus ensuring robustness in the model’s predictions. SVMs can be extended to handle non-linear classification by using kernel functions that map the input data into higher-dimensional spaces. Popular kernels include the radial basis function (RBF), polynomial, and sigmoid kernels. SVMs are especially useful in high-dimensional spaces and perform well in classification tasks with a clear margin of separation.

The **k-Nearest Neighbors (k-NN)** algorithm is an instance-based learning technique that makes predictions based on the proximity of a given test point to the labeled instances in the training set. For classification, the majority label among the nearest neighbors determines the predicted class, while for regression, the average value is often used. Despite its simplicity, k-NN can achieve high accuracy for certain problems, though its computational cost scales poorly with large datasets due to the need for pairwise distance calculations.

3.2.2 The Evolution of Neural Networks in Modern Supervised Learning

Deep learning methods, particularly those involving neural networks, have become increasingly popular for selecting and analyzing acoustic features in mental health research [75]. Different types of neural network architecture were used to resolve mental health problems.

Neural networks have fundamentally transformed the landscape of supervised learning, particularly in scenarios involving large-scale, high-dimensional, and unstructured data. Unlike classical machine learning models, which often rely on extensive manual feature engineering, neural networks have the capacity to automatically learn hierarchical feature representations from raw data. This makes them particularly suited for complex tasks involving image recognition, natural language processing, and time-series prediction, where latent patterns and structures can be captured more effectively.

The most popular Neural Networks architecture are as follows:

Artificial Neural Networks (ANNs) [44] are computational models inspired by the biological neural networks of the human brain. They consist of layers of interconnected nodes, or neurons, where each neuron computes a weighted sum of its inputs followed by the application of a non-linear activation function. This architecture enables ANNs to model non-linear relationships between input features and target outputs, a key advantage over linear models.

The training process of ANNs typically involves minimizing a loss function, such as cross-entropy or mean squared error, using gradient-based optimization techniques like stochastic gradient descent. The **backpropagation** algorithm, which calculates the gradient of the loss function concerning the network's weights, is critical in this optimization process. Through iterative weight updates, the network learns to approximate the desired output.

Despite their power, ANNs have certain limitations. Training deep networks (networks with many layers) can lead to issues such as vanishing or exploding gradients, which impede learning. Additionally, ANNs require large datasets to generalize effectively and are computationally expensive to train, often requiring specialized hardware such as GPUs.

Convolutional Neural Networks (CNNs) [59] are specialized types of neural networks designed for grid-like data, such as images or audio. It is particularly effective in processing audio spectrograms (visual representations of the frequency spectrum of audio signals) [55]. This type of NN can automatically learn and select relevant features from the data without the need for manual feature extraction. CNNs employ convolutional layers, which apply filters to local regions of the input, allowing the network to automatically learn spatial hierarchies of features, from low-level edges and textures to high-level structures like objects. This architecture is particularly effective for image recognition tasks, where local spatial relationships are critical for accurate classification.

CNNs are characterized by their ability to capture translation-invariant features, meaning they can detect patterns regardless of their position in the input space. Pooling layers, which reduce the dimensionality of feature maps while retaining important information, further enhance the efficiency of CNNs.

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) [32]: These networks are designed to handle sequential data, making them well-suited for speech recognition [33] where the temporal order of the input is essential for accurate predictions. It can automatically identify and select features related to speech patterns that indicate mental health conditions. RNNs maintain an internal state, or memory, which allows them to model dependencies across time steps. However, traditional RNNs struggle with long-term dependencies due to the vanishing gradient problem. LSTM networks address this issue by incorporating memory cells and gating mechanisms (input, forget, and output gates) that regulate information flow, enabling the retention of relevant information over extended sequences.

3.3 Semi-supervised methods for Medical Sensor Data respecting uncertainty of data

Supervised machine learning (SL) algorithms rely on labeled data, which often makes the process of labeling both labor-intensive and prone to errors. In many practical applications, obtaining labeled data is either unfeasible or limited to a few examples. In contrast, unsupervised learning techniques (UL), such as clustering algorithms, address these challenges by extracting patterns and knowledge from unlabeled data to build predictive models. However, in the absence of labeled information that reflects the underlying data distribution, clustering methods can yield insufficient results, potentially creating groupings that contain instances from different classes. Semi-supervised learning (SSL) algorithms using a combination of both labeled and unlabeled data [13] (see Figure 8).

We adopt the core principle of semi-supervised learning to enhance unsuper-

vised methods, such as clustering, by incorporating a partial supervision mechanism. This mechanism provides meaningful guidance during the process of knowledge discovery from data. Therefore, the objective of semi-supervised clustering is to identify meaningful semantic categories by grouping similar data points, leveraging domain knowledge in the form of supervised information provided explicitly by a domain expert [31].

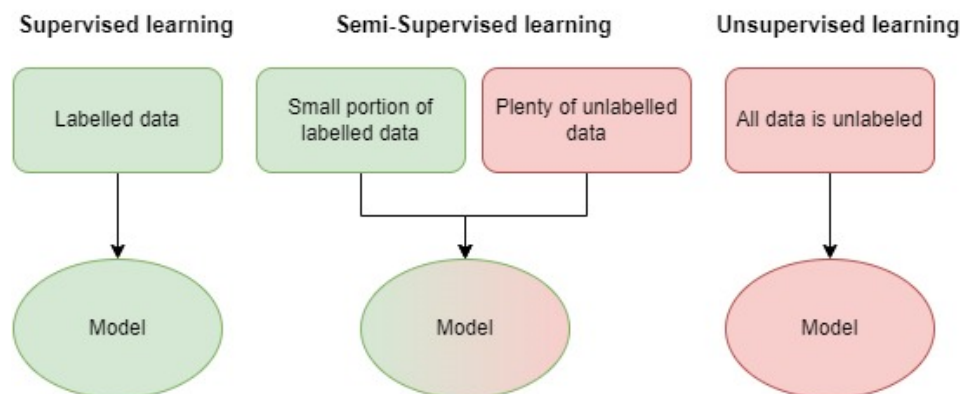


Figure 8: Semi-supervised learning emplacement

Although numerous semi-supervised algorithms have been proposed in the literature, to the best of my knowledge, no comprehensive evaluation has been undertaken to assess their effectiveness in the context of sensor-based mental health monitoring. This domain is characterized by the coexistence of labeled and unlabeled data, accompanied by inherent uncertainty.

In most related studies (e.g., [19, 22, 6, 34, 25]), predicting bipolar disorder episodes is typically addressed as a supervised learning task. To overcome challenges related to uncertainty in patients' states and the scarcity of labeled data, unsupervised learning techniques were introduced in [42] and [51]. These approaches leveraged the entirety of the available dataset rather than limiting the analysis to data from narrow time windows surrounding psychiatric assessments.

In this dissertation, we propose the SCVA method (described in Section 6 to address the challenge of evaluation clusters agreement using uncertain medical assignments. For that study firstly we use one of the popular Unsupervised Algorithm called Self-Organizing Maps.

Self-Organizing Maps (SOM) also known as Kohonen's Network were developed by Teuvo Kohonen [57] in the 1980s. It is a type of artificial neural network used for **unsupervised learning**. They are particularly useful for visualizing and interpreting high-dimensional data by projecting it onto a lower-dimensional, typically two-dimensional, grid. SOMs consist of a lattice of neurons, each associated

with a weight vector of the same dimension as the input data. During training, SOMs employ a competitive learning approach where input vectors are presented to the network, and the neuron with the weight vector most similar to the input vector, known as the Best Matching Unit (BMU), is identified. The weights of the BMU and its neighboring neurons are then adjusted to become more like the input vector, promoting topological ordering and preserving the input space structure. One of the main strengths of SOMs is their ability to preserve the topological relationships of the input data. Similar data points are mapped to nearby locations on the map, maintaining the inherent structure of the data.

3.4 The role of explainability in mental health context

In recent years, the intersection of artificial intelligence (AI) and mental health has garnered significant attention. AI-driven systems are increasingly employed to support mental health diagnostics, treatment plans, and ongoing patient monitoring. However, one of the most critical and often underexplored aspects of AI in mental health is explainability. The ability of AI systems to provide clear, understandable reasoning behind their decisions and recommendations is paramount in this context.

Mental health is a deeply personal and complex domain where trust between patients and healthcare providers is essential [3]. The introduction of AI adds an additional layer of complexity, as these systems often function as “black boxes,” making decisions that are not easily understood by humans. Explainable AI (XAI) [83] addresses this challenge by providing clinicians with insights into how AI systems arrive at their recommendations. This transparency allows healthcare professionals to critically assess AI suggestions and make more informed decisions, thereby enhancing trust and effectiveness in mental health care.

As the importance of explainability in AI systems becomes more evident, several frameworks and tools have been developed to make machine learning models more interpretable and transparent. These tools are crucial in domains like mental health, where decisions based on AI recommendations must be easily understandable and justifiable to both clinicians and patients. This section explores some of the most widely used frameworks for Explainable AI, including SHAP and LIME [74], and their relevance to mental health applications.

SHAP (SHapley Additive exPlanations) [63] is one of the most widely adopted frameworks for explainability in AI. Based on cooperative game theory, SHAP provides explanations for machine learning models by assigning each feature an importance value that reflects its contribution to the model’s prediction. The strength of SHAP lies in its ability to offer consistent and theoretically sound explanations that are both local (specific to individual predictions) and global (reflective of the overall model behavior).

Explainable AI methods can be broadly categorized into two types:

- **Global Explainability** - These methods provide an overall view of how the model makes predictions, explaining the general patterns and behaviors of the model across the entire dataset. For example, feature importance rankings can offer insights into which features most influence the model’s predictions.
- **Local Explainability** - These methods focus on explaining individual predictions, i.e., why the model made a specific prediction for a particular input. Local interpretability is particularly important when individual decisions need justification.

SHAP begins by determining a baseline prediction, typically calculated as the average output of the model across the entire dataset. This baseline serves as a reference point to measure how much each feature contributes to deviating from the average prediction.

For each instance in the dataset, SHAP evaluates how each feature affects the model’s prediction by forming different subsets, or coalitions, of features. Each subset represents a different combination of features, where the feature in question is either included or excluded from the coalition.

SHAP calculates the marginal contribution of a feature by comparing the model’s prediction when the feature is included in the coalition versus when it is excluded. This process is repeated for all possible feature subsets to account for how a feature’s contribution changes in different contexts.

The Shapley value for a particular feature is computed by averaging its marginal contributions across all possible feature coalitions. This ensures that the importance of a feature is fairly evaluated in combination with other features. The Shapley value represents the average contribution of the feature to the model’s output for a given instance.

SHAP constructs an additive explanation model for each prediction using the following formula:

$$\hat{y}(x) = \phi_0 + \sum_{i=1}^M \phi_i \quad (3)$$

Where:

- $\hat{y}(x)$ is the predicted value for instance x .
- ϕ_0 is the baseline prediction, usually the average output of the model.
- ϕ_i is the Shapley value for feature i , representing the contribution of feature i to the deviation from the baseline.

4 Feature selection

The following section focuses on one of the key stages in the machine learning model development process – selecting the appropriate input features that have the most significant impact on prediction quality and result interpretability. Within the broader field of computational intelligence, feature selection occupies a fundamental role, as it directly influences the efficacy of various intelligent systems by optimizing the data input to algorithms. By carefully choosing relevant features, practitioners can enhance model performance, reduce computational complexity, and improve the system’s ability to learn and adapt from data.

In the context of mental health monitoring, where the set of features may include data from various sources such as mobile recordings therefore voice fluctuations, activity or biometric data, the proper choice of features becomes particularly important. Feature selection serves several critical functions.

In that section, we introduced a new method for selecting important acoustic features that take into account the uncertainty of patient assignments.

In this Section, we discuss the uncertainty aware approach for feature selection. The primary goal of the Uncertainty-aware advanced Feature Selection (UAAFS) approach is to improve predictive performance due to adequate consideration of imprecision related to the sensor data and psychiatric assessments during the feature engineering and selection stage. In the context of remote monitoring for bipolar disorder, it is common to extend the psychiatrist’s evaluation to a specific time frame surrounding the visit, known as the ground truth period.

Beside that in subsection 4.3 there is another approach that present problem of features selection preserving cost of features acquisition.

4.1 The Role of Feature Selection in Computational Intelligence

Feature selection is essential for building models that are not only accurate and efficient but also interpretable and generalizable. It optimizes the balance between model complexity and performance, making it a key step in the machine learning pipeline. The following advantages were mentioned by the authors [77] as the most important advantages of feature selection:

- to avoid overfitting and improve model performance
- to provide faster and more cost-effective models
- to gain a deeper insight into the underlying processes that generated the data

According to [36], we could divide the area of feature selection into wrappers, filters, and embedded methods presented in Figure 9.

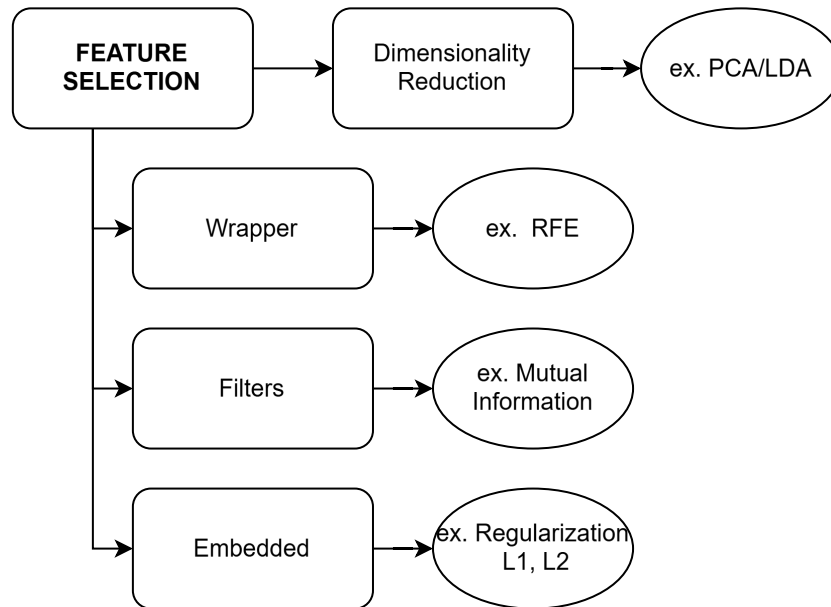


Figure 9: Division of Feature Selection methods

Wrappers utilize the learning machine of interest as a black box to score subsets of variables according to their predictive power. The well-known technique is RFE, which is described in that section. Filters select subsets of variables as a pre-processing step, independently of the chosen predictor. These methods use statistical techniques to score each feature independently. Some examples are the Pearson Correlation Coefficient and Mutual Information. Embedded methods perform variable selection in the training process and are usually specific to given learning machines. These methods perform feature selection during the model training process. The model itself decides which features are most important. Best examples are Lasso (L1 Regularization), which encourages sparse models by penalizing the absolute magnitude of feature coefficients, forcing some to become zero and Ridge (L2 Regularization), that Penalizes large coefficients to prevent overfitting but does not force coefficients to zero. Moreover, the author distinguishes another field of feature selection named Feature Construction and Space Dimensionality Reduction. In certain applications, reducing data dimensionality by selecting a subset of the original variables can be advantageous due to factors such as the cost associated with acquiring, storing, and processing measurements. However, when these constraints are not significant, alternative dimensionality reduction methods, such as feature extraction or transformation techniques, should also be considered

to optimize computational efficiency and model performance. The best-known example is PCA (Principal Component Analysis), which is a statistical technique used to reduce the dimensionality of large datasets while preserving as much variance as possible. It achieves this by transforming the original variables into a new set of orthogonal components called principal components. These components are ranked in order of the amount of variance they explain, with the first few typically capturing the majority of the information in the data.

Several challenges are encountered in this process. We possess a substantial volume of acoustic data, but only a limited subset of features is responsible for detecting significant characteristic changes in mania [27] and depression [87], [69], [76]. Moreover, from the entire available dataset, only a small proportion of data points are labeled, necessitating feature selection to be performed on this labeled subset. Our objective is to identify a few critically important acoustic features and determine predictors that characterize mania and depression while accounting for the uncertainty inherent in the labels. Additionally, we aim to compare these predictors across different target groups to ascertain whether they exhibit global patterns (common across all patients) or individual patterns (specific to each patient).

4.2 Uncertainty Aware Advanced Feature Selection

The Uncertainty Aware Advanced Feature Selection method (UAAFS) introduces information about label uncertainty to select the most relevant features in the data set. Uncertainty aware advanced feature selection method (UAAFS) consists of several information about label uncertainty to the selection of the most relevant features in the data set.

4.2.1 Method overview

Combining the described label confidence, Recursive Feature Elimination (RFE) - described in Section 4.2.2, and extreme gradient boosting (XGBoost) - described in Section 4.2.3, we are receiving Uncertainty aware advanced feature selection (UAAFS) method that could model the feature selection problem using acoustic data and its uncertainty information about psychiatric assignment. This approach aims to address problem of selection important features of this dissertation.

In this approach, RFE uses the feature importance scores provided by the XGBoost model to decide which features to eliminate in each iteration. XGBoost calculates these importances either through the "gain", which measures the improvement in accuracy brought by a feature to the branches it is on, or through other metrics like "cover" and "frequency". By using these importance scores, RFE can more accurately identify and eliminate the least significant features.

The process starts with training an XGBoost model on the entire set of features and computing the importance scores. Label confidence factors are joined to the entire dataset and set as weights for the XGB model. RFE then eliminates the least important feature(s), and the model is retrained on the reduced feature set. This cycle repeats until the desired number of features is reached or until no further improvement is observed in the model's performance. That algorithm is described in Algorithm 1.

Algorithm 1 Uncertainty Aware Advanced Feature Selection algorithm

```

1: Input:  $X$ : Feature vector  $W_{rp}$ 
2:  $y$ : Target variable ( $m \times 1$ )
3: model: The machine learning model to use
4:  $k$ : Number of features to select
5: Output: selected_features: List of  $k$  selected features
6: Initialize selected_features as an empty list
7: Initialize  $n$  as the number of features in  $X$ 
8: while  $n > k$  do
9:   Train the XGB model on  $X$  and  $y$  with label confidence as a weights
10:  Compute the feature importance for each feature
11:  Identify the feature with the lowest importance
12:  Remove the feature with the lowest importance from  $X$ 
13:  Decrement  $n$  by 1
14: end while
15: selected_features = the remaining features in  $X$ 
16: return selected_features

```

In XGBoost, instance weights provide a mechanism to assign different importance to various samples in the training dataset. This feature is particularly useful in diverse datasets where some instances are more significant than others, such as in cases with imbalanced data or in domains where newer instances are more relevant. During the model training phase, these weights influence the calculation of the loss function, effectively making some instances count more in the computation of gradients and updates to the model. Higher weights mean that the model will focus more on reducing errors for those instances, making them pivotal in shaping the model's understanding and predictions. This capability allows XGBoost to adapt to specific requirements of different training scenarios, enhancing its versatility and applicability in practical situations where data discrepancies or imbalances might otherwise skew the model's performance.

For classification, the loss function (6) is replaced with the softmax loss function and is defined as:

$$L(y, \hat{y}) = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^K y_{ij} \log(\hat{y}_{ij}) \quad (4)$$

where:

y : True label (one-hot encoded)

$\hat{y}$: Predicted probability for each class

n : Number of samples

K : Number of classes

To sum up, medical labels are often subjective opinions that can vary depending on the day or the individual. Accounting for this uncertainty associated with each label at the initial stages of modeling, particularly during feature selection, allows for a more accurate alignment of the model with the problem being addressed.

The UAAFS method leverages an efficient XGBoost model, which has been integrated into the RFE algorithm framework for feature selection. It also incorporates label uncertainty into the process.

4.2.2 Recursive Feature Elimination

The UAAFS method is based on Recursive Feature Elimination (RFE) algorithm [37]. RFE is a powerful feature selection method that enhances the performance of machine learning models by systematically removing the least significant features. This technique operates within a backward selection process where it constructs a model on an initial set of features and evaluates their importance, typically using a coefficient in linear models or feature importance scores in ensemble methods. RFE iteratively refines the feature set by eliminating the weakest feature at each step, which is determined based on the model's performance criteria. This process continues until a specified number of features is reached or until further removal no longer improves the model's performance. The effectiveness of RFE lies in its ability to reduce overfitting and improve model generalizability by retaining only the most relevant features. It is beneficial in scenarios where the dataset contains many features, some of which may be redundant or irrelevant, thereby streamlining the model to focus on the most impactful predictors.

4.2.3 Selected classification model: XGBoost

The classification model there were selected in the UAAFS method is XGBoost, or Extreme Gradient Boosting, introduced by Tianqi Chen [14]. It is a machine learning algorithm that represents a highly efficient implementation of gradient boosted decision trees. Gradient boosting [29] is a technique utilized for both regression and classification tasks, where models are constructed sequentially. Each new model is trained to correct the errors made by its predecessors, resulting in a final model that is a weighted sum of all the individual models. It is particularly noted for its capability to handle diverse data types, various data distributions, and missing values, making it a robust choice for a wide range of machine learning applications.

In XGBoost, the objective function is crucial in guiding the training process. It is composed of two main components: the loss function $L(\theta)$ and the regularization term $\Omega(f)$. The primary goal of the objective function is to measure and minimize the overall error of the model while simultaneously preventing overfitting by penalizing model complexity.

$$\text{Obj}(\theta) = L(\theta) + \Omega(f) \quad (5)$$

The loss function, $L(\theta)$, quantifies the discrepancy between the predicted values and the actual target values. The loss functions often used for classification tasks are log loss and hinge loss. The mathematical representation of the loss function for n training examples is as follows:

$$L(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) \quad (6)$$

In this equation, y_i represents the actual target value, $\hat{y}_i$ denotes the predicted value, and l is the specific loss function applied.

To address the issue of overfitting, XGBoost incorporates a regularization term, $\Omega(f)$, into the objective function. The regularization term penalizes the complexity of the model, thus encouraging the learning of simpler models that better generalize unseen data. The regularization term can be expressed as:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (7)$$

Here, T indicates the number of leaves in the tree, w_j represents the weight associated with leaf j , γ is a regularization parameter that controls the number of leaves, and λ is a parameter that applies L2 regularization on leaf weights.

This objective function balances the trade-off between the model's accuracy and complexity. By minimizing this objective function, XGBoost ensures that the model fits the training data well and remains robust and generalizable to new data.

XGBoost constructs its model by sequentially adding decision trees, each designed to correct the errors made by the previous trees. This additive approach allows the model to improve incrementally, refining predictions with each new tree added.

One of the critical concepts in XGBoost is gain, which measures the improvement in accuracy brought by a feature or split to the model. When constructing trees, XGBoost evaluates all possible splits and selects the one that maximizes the gain. This approach ensures that each split meaningfully improves the model's predictive power. Gain is a vital criterion in split finding, as it directly influences the tree's structure and the importance assigned to each feature.

Weights play a significant role in Gradient Boosted Decision Trees models construction process. Each training instance can be assigned a weight, reflecting its importance. These weights influence the calculation of the gradients and the gain during the split-finding process. By assigning higher weights to more critical or underrepresented instances, the model can be guided to focus more on these important cases. This is particularly useful in scenarios with imbalanced datasets, where certain classes or instances require more attention to achieve accurate predictions.

In addition to maximizing gain, XGBoost uses tree pruning to eliminate splits that do not significantly reduce error. This pruning process starts by growing trees to their maximum depth and then removing branches that add little to the model's performance. By focusing on splits that offer substantial gain, pruning helps enhance efficiency and prevents overfitting, where the model becomes overly complex and performs poorly on new, unseen data.

To further control overfitting, XGBoost integrates regularization methods. It includes both L1 (Lasso) and L2 (Ridge) regularization, which penalizes the complexity of the model. L1 regularization encourages sparsity by penalizing the absolute values of leaf weights, while L2 regularization penalizes the squared values of the weights, helping to maintain simplicity in the model. Additionally, XGBoost utilizes a learning rate, also known as shrinkage, which scales the contribution of each tree. This gradual learning process helps to avoid overfitting by moderating the pace of learning.

Moreover, XGBoost can employ column subsampling, where only a random subset of features is used to build each tree. This adds an element of randomness to the model, further reducing the risk of overfitting and enhancing generalization.

4.3 Cost Constrained Feature Selection

In this section, I would like to present an entirely different perspective on feature selection using the Cost-Constrained Feature Selection method (CCFS), which tries to resolve the problem of selecting important features.

This approach is widely used in the different fields of medicine, where a specific cost is assigned to medical examinations in a given currency. By simplifying this problem and not considering the applicable currency, we can use prime numbers to imagine the scale of the cost problem. Assuming that blood tests are usually the cheapest form of medical examination, we could assign value q as a predictor for specific diseases. Another important factor could be described with Ultrasonography (USG), computed tomography (CT), and Magnetic Resonance Imaging (MRI), that are naturally more expensive and successively are given the following weights ($3q$, $5q$, $10q$) - visual representation available in Figure 10. Sometimes, the choice of a significant examination and the features it offers are unobvious. Therefore, the CCFS method could be helpful. It can identify which features provide the most valuable information for a given problem, considering the assumed cost of the examination.

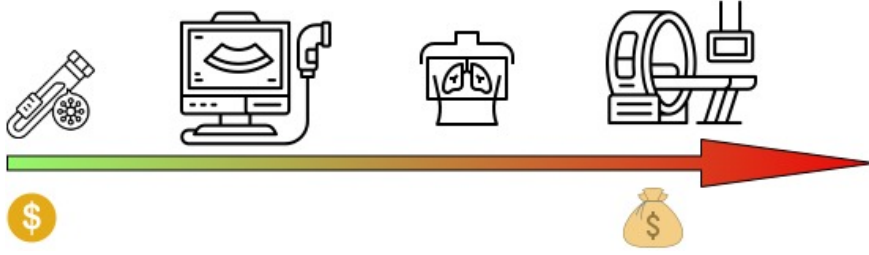


Figure 10: Illustrated cost of obtaining features

In the context of sensor data that this dissertation is about, we are using not just acoustic features but also some behavioral features that were taken into account (such as the number of incoming/outgoing calls per day or the average length of characters in daily text messages). All of the features were represented by daily aggregated values using vector W_{rp} and it related labels equally confident L_{dp} . That approach is particularly described in authors paper [50].

First of all CCFS introduce cost function that represents different type of possessing data. In our case, two different costs were taken into account. The first cost equal to 1 is assigned to behavioral data - as basic call statistics. Next, the cost equal to 5 (author conducted the comparison and assume that this value is acceptable) is represented by acoustics features a .

The idea of CCFS is based on a greedy forward selection algorithm using information theory [82].

Using approach A2 of data representation mentioned in Section 1.3 we could assume that :

$a_1, \dots, a_p$ be the features, and L be a multiclass label, and the rest of the required coefficients remain unchanged.

Next, using an equation from the original paper [50] with : In each step of the algorithm, we add the index of the candidate feature a_k to the set of features already selected in the previous steps $S \leftarrow S \cup \{k^*\}$ by maximizing the following Equation:

$$k^* = \arg \max_{k \in F \setminus S} [J(X_k, Y|X_S) - \lambda c(k)], \quad (8)$$

where $J(X_k, Y|X_S)$ measures the informativeness of the added feature X_k in the context of already selected features and $\lambda c(k)$ is the penalty for the cost of the added feature. The general idea behind Equation (8) is that it allows us to find the trade-off between the relevance of the feature and its costs and parameters, such as λ that controls the balance. The optimal parameter λ value can be calculated by minimizing the loss function with the cross-validation on the training set.

The main goal of the experiments was to compare the cost-constrained algorithm for feature selection and to compare it with its traditional counterpart.

The current study indicates that implementing only a small part of acoustic features (which are time/cost related for researchers) could be sufficient to receive first-worth results.

5 Explainable supervised learning in mental health monitoring

Supervised learning is a crucial component of computational intelligence, serving as a primary method by which machines infer relationships between inputs and outputs using labeled datasets. It plays a key role in constructing predictive models that support automated decision-making, classification, and regression tasks.

Currently, in the rapidly evolving world of artificial intelligence technology, classification results play a crucial role in decision-making across various sectors, including healthcare and many others. However, equally important—and often underestimated—are Explainable AI (XAI) methods. XAI is gaining significance because it not only provides answers but also explains how these answers were achieved, which is essential for a comprehensive understanding and trust in AI systems. This issue is particularly pertinent with the rise of currently popular neural networks, which, despite their impressive performance, often operate as “black boxes” making their decision-making processes opaque and difficult to interpret.

In that section beside fundamental classification of BD state based on mobile recordings I would like to present complex model that satisfies the needs for explaining relations between attributes, symptoms, and states maintaining label extrapolation.

5.1 Classification of Bipolar Disorder states with selected Neural Networks architecture

As mentioned in Section 2.2, distinguishing the state of a patient with bipolar disorder is one of the key tasks that can be helpful in the remote monitoring of patients using their voice recordings. Nowadays, the popularity of neural networks is immense, so we have prepared a comparison of three elementary neural network architectures to evaluate how they handle this data in patient classification.

The following neural network architectures were selected for the study.

1. **Artificial Neural Networks (ANN)** represent the most fundamental form of neural network architectures.

The network starts with an Input layer containing 64 neurons and using the ReLU activation function. This layer includes L2 regularization with a factor of 0.01 to penalize large weights. Following this, a Dropout layer with a 50% rate randomly deactivates neurons during training, preventing overfitting. The second Dense layer consists of 32 neurons with ReLU activation and the same L2 regularization applied. It is succeeded by another Dropout layer with a 40% rate. The network concludes with a Dense output layer that uses the softmax activation function to generate probability distributions over the target classes. Visualization of its architecture is presented in Figure 11

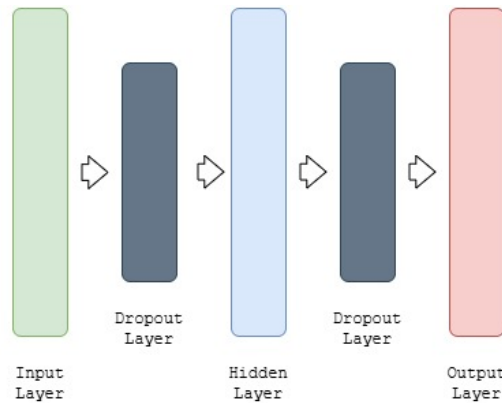


Figure 11: Proposed basic ANN architecture for BD states classification

Dropout - is a regularization technique employed in neural networks to mitigate overfitting, thereby enhancing the model's generalization capabilities. It operates by randomly deactivating a subset of neurons during each training iteration. Specifically, during the forward pass, each neuron is independently excluded

typically ranging between 20% and 50%. This stochastic omission forces the network to develop redundant representations and prevents co-adaptation of neurons, ensuring that the model does not become overly reliant on specific pathways for making predictions.

2. **Convolutional Neural Networks (CNN)** are specialized neural networks designed primarily for processing data with a grid-like topology, such as images or signals. The network starts with a Conv1D layer of 64 filters and a kernel size of 3, activated by ReLU and regularized with L2. This is followed by MaxPooling1D and a Dropout layer with a 40% rate to downsample and regularize the data. A second Conv1D layer with 128 filters captures more complex features, followed by another MaxPooling1D and a 30% Dropout layer. The extracted features are then flattened with Flatten layer and passed through a Dense layer with 64 neurons and ReLU activation, further regularized by L2. A final Dropout layer with a 20% rate precedes the Dense output layer, which uses softmax activation to produce probability distributions over the target classes. Visualization of its architecture is presented in Figure 12

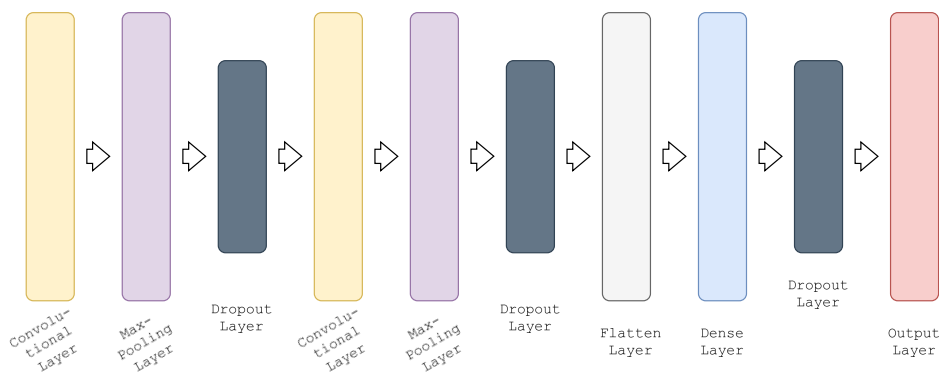


Figure 12: Proposed basic CNN architecture for BD states classification

Convolutional layer - is a fundamental component of Convolutional Neural Networks (CNNs), widely used for processing recordings such as audio signals. Its primary role is automatically detecting and learning important audio data features. In a convolutional layer, small matrices called filters or kernels move over the input data represented as the recordings. Each filter is designed to identify specific characteristics in the audio. As a filter slides across the input, it performs element-wise multiplication and sums the results to produce a new output called a feature map. This feature map indicates where particular audio features are present in the recording. Activation functions, such as ReLU (Rectified Linear Unit), are

applied after the convolution operation to introduce non-linearity. This allows the network to learn more complex and abstract patterns in the audio data.

MaxPooling - is a simple yet important part of CNNs used for processing recordings like audio signals. Its main job is to reduce the data size while keeping the most important information. The max pooling layer looks at small sections of the data called windows. In each window, it picks the highest value and ignores the rest. This process is repeated across the entire data, making the output smaller and more straightforward.

3. **Long Short-Term Memory Networks (LSTM)** that were developed as an advanced variant designed to better capture long-term dependencies within sequential data. The network begins with a 64-unit LSTM layer configured to return sequences, both kernel and recurrent weights regularized with L2 to prevent overfitting. This is followed by a Dropout layer with a 30% rate to enhance regularization further. A second 64-unit LSTM layer processes the sequences without returning them, also employing L2 regularization. The extracted features are then passed through a Dense layer with 64 neurons and ReLU activation, maintaining simplicity through L2 regularization. Finally, a Dense output layer with softmax activation generates probability distributions over target classes represented BD state. Visualization of its architecture is presented in Figure 13.

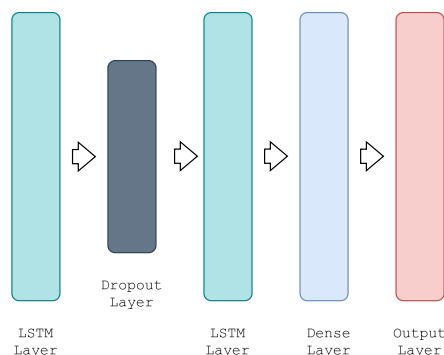


Figure 13: Proposed basic LSTM architecture for BD states classification

LSTM layer - is an essential component in neural networks, especially when working with recordings like audio signals. Its main purpose is understanding and remembering patterns over time, making it ideal for processing sequential data. LSTM layers have special units called memory cells that can store information for long periods. These cells help the network remember important details from earlier in the recording. To manage what information is kept or discarded, LSTM layers

use gates. The input gate decides what new information to add to the memory, the forget gate chooses what information to remove, and the output gate determines what information to send to the next layer. This combination allows LSTM layers to control the flow of information effectively, ensuring that relevant details are retained while unnecessary data is filtered out.

5.2 Explainable Compositional MLP approach for the disease and its symptoms

In [48], we proposed a novel approach called PLENARY (illustrated in Figure 14), which consists of three modules:

- Creation of a compositional classification model via supervised learning based on a two-level hierarchy of labels associated with data
- Explanation of the predictive model's outcomes using SHAP
- Creation of linguistic summaries on global model explanations using fuzzy quantified sentences

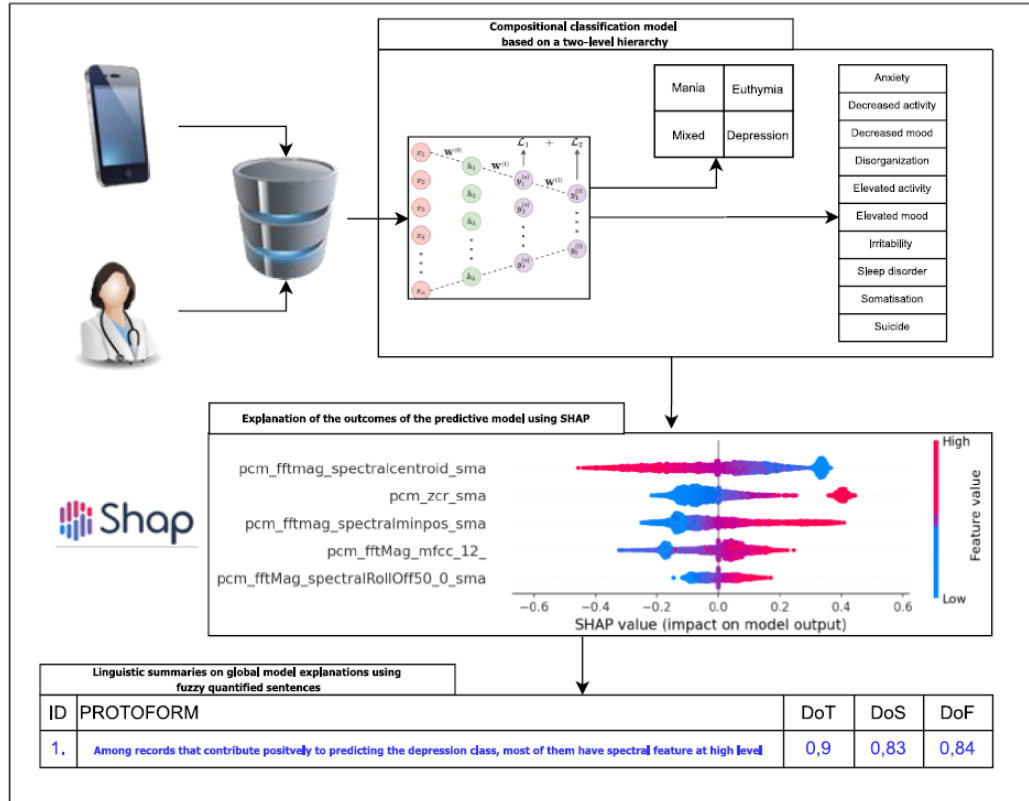


Figure 14: A three-part approach called PLENARY [48] that represents compositional classification model, explanation of the predictive model's and linguistics summaries

The third component of PLENARY involves creating linguistic summaries based on SHAP values. These summaries provide further explanation of significant parameters that impact the classification of the state and symptoms of bipolar

disorder. PLENARY [48] aimed to leverage XAI system outputs to provide non-technicians, such as medical experts, with linguistic summaries that explain the relationship between feature importance and model outputs. We adapted proto-forms [45] and fuzzy set computations to link groups of attributes with global explanations.

In this dissertation, we address the problem of labels uncertainty and introduce that information to different training samples and propose **XCMLPCF** (**EX**plainable **CO**mpositional **ML**ayer **PC**eptron with **CF** Confidence **F**actor). That study does not evaluate linguistic summaries. Instead, we are focusing on label uncertainty, therefore the current approach. The final flow of that approach is presented in Figure 15.

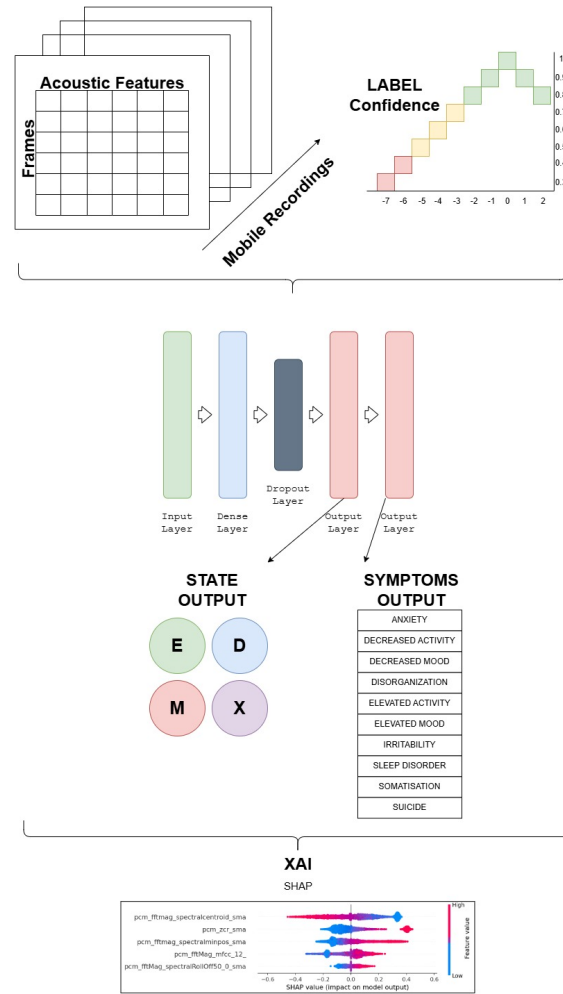


Figure 15: The workflow of Explainable Compositional Multi-Layer Perceptron with Confidence factor (**XCMLPCF**)

The proposed approach consists of two modules:

- Creation of a compositional classification model including confidence factor for labeling based on a two-level hierarchy of labels associated with data;
- Explanation of the predictive model’s outcomes using SHAP;

The first module consists of *multi-output* sequential and compositional MLP trained to simultaneously predict two levels of labels (symptoms and mental states in our case study) from the same data. Figure 16 illustrates the model architecture. The network architecture combines two output layers to predict both targets $\mathbf{y}^{(s)}$ and $\mathbf{y}^{(t)}$ simultaneously. We denote by $\hat{\mathbf{Y}}^{(s)}$ and $\hat{\mathbf{Y}}^{(t)}$ the network outputs approximating these targets for all n training examples.

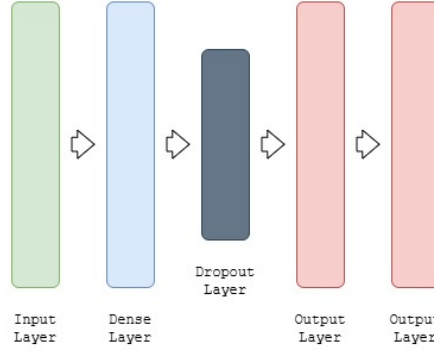


Figure 16: Architectural diagram of the multi-task neural network in PLENARY [48].

Formally, the final output of the model is:

$$\hat{\mathbf{Y}}^{(t)} = \text{softmax} \left(\hat{\mathbf{Y}}^{(s)} \mathbf{W}^{(2)} + \mathbf{b}^{(2)} \right), \quad (9)$$

where $\mathbf{W}^{(2)} \in \mathbb{R}^{s \times t}$ and $\mathbf{b}^{(2)} \in \mathbb{R}^t$ are the weights and biases of the final output layer. The intermediate output $\hat{\mathbf{Y}}^{(s)} \in \mathbb{R}^{n \times s}$ is defined as:

$$\hat{\mathbf{Y}}^{(s)} = \mathbf{H} \mathbf{W}^{(1)} + \mathbf{b}^{(1)}, \quad (10)$$

with $\mathbf{W}^{(1)} \in \mathbb{R}^{h \times s}$ and $\mathbf{b}^{(1)} \in \mathbb{R}^s$ being the weights and biases of the intermediate output layer, and $\mathbf{H} \in \mathbb{R}^{n \times h}$ the output of a hidden layer with h units:

$$\mathbf{H} = \text{ReLU} \left(\mathbf{X} \mathbf{W}^{(0)} + \mathbf{b}^{(0)} \right), \quad (11)$$

where $\mathbf{W}^{(0)} \in \mathbb{R}^{d \times h}$ and $\mathbf{b}^{(0)} \in \mathbb{R}^h$ are the weights and biases of the hidden layer. To reduce overfitting, a dropout layer with rate $p = 0.2$ follows the hidden layer. We

use the ReLU activation function in the hidden layer and the softmax activation function in the final layer for multi-class classification.

The network minimizes the cross-entropy loss function for the final output:

$$\mathcal{L}_2 = \mathcal{H}(\mathbf{y}^{(t)}, \hat{\mathbf{y}}^{(t)}) = - \sum_{j=1}^t y_j^{(t)} \log \hat{y}_j^{(t)}, \quad (12)$$

where $\hat{\mathbf{y}}^{(t)}$ are the predicted class probabilities, and $\mathbf{y}^{(t)}$ is the one-hot encoded ground truth vector of length t .

For the intermediate output, the activation function depends on the nature of the mid-level annotations. Suppose these annotations are numeric values (e.g., symptoms represented as ordinal values in our case study). In that case, no activation function is used, and the layer performs multi-output regression with the mean absolute error loss:

$$\mathcal{L}_1 = \text{MAE}(\mathbf{y}^{(s)}, \hat{\mathbf{y}}^{(s)}) = |\mathbf{y}^{(s)} - \hat{\mathbf{y}}^{(s)}|, \quad (13)$$

where $\mathbf{y}^{(s)}$ and $\hat{\mathbf{y}}^{(s)}$ are the ground truth and predicted values for each training example, respectively. If mid-level annotations are class labels, then $\mathcal{L}_1$ is the cross-entropy loss as in $\mathcal{L}_2$. The network minimizes the composite loss function:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2. \quad (14)$$

Next, we use sample weights for the loss functions during training process. Specifically, each sample was determined based on the step function described in Section 2.3, and the weight w_c was computed as:

$$w_c = \begin{cases} 0.3 & : \text{time_point} = -7 \\ 0.4 & : \text{time_point} = -6 \\ 0.5 & : \text{time_point} = -5 \\ 0.6 & : \text{time_point} = -4 \\ 0.7 & : \text{time_point} = -3 \\ 0.8 & : \text{time_point} = -2 \\ 0.9 & : \text{time_point} = -1 \\ 1 & : \text{time_point} = 0 \\ 0.9 & : \text{time_point} = 1 \\ 0.8 & : \text{time_point} = 2 \end{cases} \quad (15)$$

Therefore, the weighted loss function becomes:

$$\mathcal{L}_{\text{weighted}} = \frac{1}{N} \sum_{i=1}^N w_i (\mathcal{L}_{1,i} + \mathcal{L}_{2,i}), \quad (16)$$

where w_i is the weight assigned to the i -th training sample, and $\mathcal{L}_{1,i}$ and $\mathcal{L}_{2,i}$ are the losses for the intermediate and final outputs for that sample, respectively. This weighting mechanism ensures that the model gives appropriate attention to samples from different classes or with varying levels of label confidence during training.

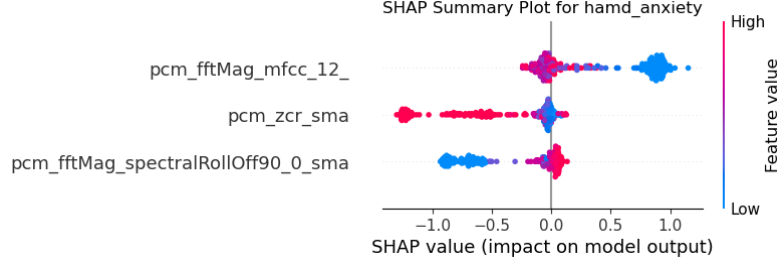


Figure 17: An example of SHAP global explanation that present an impact of three acoustic features for hamd_anxiety classification

In practice, the predictions from the intermediate output layer are used directly as new features for the final classification. Since the two levels of labels are highly interconnected, this multi-task strategy aims to improve the accuracy and consistency of the explanations for both outputs. Optimizing the first output serves as an auxiliary task to support the primary classification.

An illustrative example of SHAP analysis in a global explanation (summary) plot is shown in Figure 17. Each point represents a data instance, with color indicating the feature value range. SHAP displays the model output for a given class (e.g., depression diagnosis) or symptom (e.g., anxiety) as an inverted pyramid of the most contributing features to that class. Features with the highest contributions (i.e., highest absolute SHAP values) are placed at the top, with lower contributions arranged below. Positive contributions towards the class are shown on the positive side of the X-axis (representing positive SHAP values), while negative contributions are on the negative side.

6 Semi-supervised Clustering

Semi-supervised clustering combines the strengths of supervised and unsupervised methods, making it an especially valuable tool in mental health research where comprehensive and accurate labels are often unavailable. By leveraging both limited labeled data and abundant unlabeled data, this approach uncovers intrinsic patterns and latent groupings within complex datasets—patterns that may correspond to different symptom clusters, behavioral trends, or stages of mental health conditions.

In mental health research, fully supervised approaches are hardly possible. Patients are not continuously monitored, and mental health states can fluctuate rapidly, which limits the feasibility of obtaining detailed, ongoing assessments. The semi-supervised approach addresses these issues by incorporating available label information to guide the clustering process, thus enhancing the model’s robustness and interpretability. It enables the detection of underlying patterns even when the data is fragmented or uncertain.

A novel validation method has been developed to ensure that the clustering results remain consistent over time, particularly around key events such as medical visits. This method involves clustering recordings using various configurations of the Self-Organizing Map algorithm, which helps to differentiate between potential mental health states without directly assigning specific interpretations to the clusters. Instead, the emphasis is placed on the stability and consistency of the clustering outcomes. By comparing the frequency distributions of clusters across different periods, the validation measure assesses the agreement between clusters, providing a robust means of evaluating the semi-supervised model.

The combination of semi-supervised clustering with a focused validation strategy contributes to a deeper and more reliable understanding of mental health patterns, ultimately supporting more informed clinical decision-making and enhanced patient care.

Based on the aforementioned arguments, in the following section, I will present the Semi-supervised Cluster Validation Agreement (SCVA) method. That method is an extension of coefficient CVA presented in my first scientific paper [51]. The presented preliminary results confirm the existence of a relation between the outcome of unsupervised learning (clustering) and the psychiatric assessment. Based on that research method called SCVA were established and include uncertain nature of assessments and its presented in that section.

6.1 The Role of Unsupervised Learning in Computational Intelligence

Unsupervised learning encompasses a class of machine learning algorithms that operate without labeled data. Unlike supervised learning, which relies on input-output pairs for training, unsupervised learning explores the underlying structure of input data without predefined labels. Its primary objective is to uncover hidden patterns, relationships, and distributions within the data. The fundamental techniques in unsupervised learning include: **Clustering:** Groups similar data points based on specific criteria. The best known algorithm is *k-means*. **Dimensionality Reduction:** Reduces the number of variables to simplify models and enhance data visualization. **Association Rules:** is a technique to uncover hidden relationships between items or variables in large datasets. Mentioned examples are widely described in [40].

Moreover, UL offers a promising alternative by leveraging the vast amounts of unlabeled data available in healthcare. Its advantages include Cost-Effectiveness, which eliminates the need for expensive and laborious labeling processes. Next, it discovers novel insights, which approach identifying new patterns and structures within large and complex datasets. Therefore, UL can be applied to various types of medical data, including EHRs, genomic data, and medical imaging.

On the other hand, particularly in clustering, one of the primary challenges is validating the correctness and meaningfulness of the clusters formed. Unlike supervised learning, where performance can be directly measured against labeled outcomes, unsupervised learning lacks predefined labels or ground truth for comparison. This makes it difficult to assess whether the identified clusters truly represent meaningful patterns or are merely artifacts of the data. Without labeled data, determining the accuracy of clusters relies on indirect methods, such as internal validation metrics (like the Dunn Index, Davies-Bouldin Index, and Silhouette Index) or domain expertise. The interpretation of clusters often requires input from medical professionals to ensure that the groupings are clinically relevant and useful for decision-making.

Addressing these validation challenges requires the development of robust evaluation frameworks, collaboration between data scientists and healthcare professionals, and integrating domain knowledge to ensure that the discovered patterns are both accurate and meaningful.

6.2 The Cluster Validation Agreement

In our research, we designed a multivariable prediction model that integrates both unsupervised and supervised learning methods. The workflow of our proposed approach is illustrated in Figure 18. Data were collected during an observational study using the BDmon app installed on the smartphones of patients with bipolar disorder. The data preprocessing stage primarily involved preparing acoustic features and establishing the ground truth for psychiatric assessments. For our analyses, we considered the ground truth to encompass seven days before a clinical visit and two days afterward [34] with an equal degree of validity of the label. Next, trained and tested self-organizing maps. The resulting clusters were subsequently used as input for the random forest algorithm. Finally, we calculated the degree of agreement between clusters to assess how well the learned clusters aligned with psychiatric assessments. Due to significant variability between patients, we opted for a patient-specific model rather than a general one.

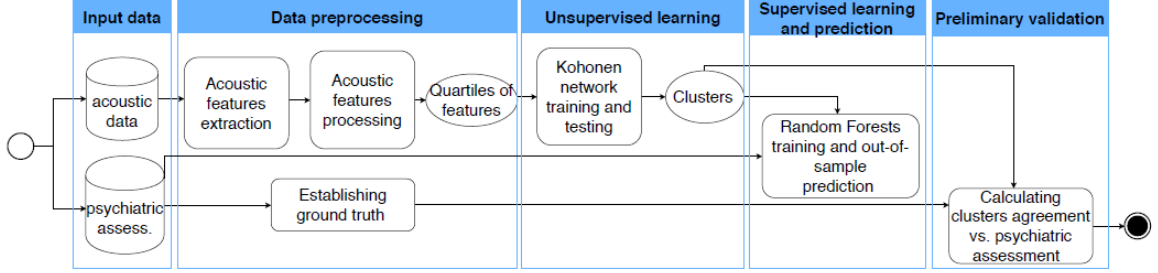


Figure 18: The workflow diagram presented in [51]

Focusing on the last step in the presented approach, let's dive into Cluster Validation Agreement (CVA) metrics.

To compare the similarity between the distributions of clusters, we first exclude days with missing data, such as those without phone calls. Next, for the days surrounding each clinical visit, we compute a vector that represents the frequency of each cluster. Let C_t denote the cluster assigned to a particular patient on day t . The frequency of the i -th cluster for the days around the visit on day t_v is given by:

$$f_{t_v, i} = \frac{\sum_{t=t_v-7}^{t_v+2} I(f_{c_t} = i)}{\sum_{t=t_v-7}^{t_v+2} I(c_t \text{ is not missing})} \quad (17)$$

where I is the indicator function taking value one if the predicate in curly brackets is true and zero otherwise, and c_t denotes a cluster which was assigned to on day t to a particular patient.

Then, we compare the distributions f_{tA} and f_{tB} for two visits A and B, with the normalized absolute difference:

$$a_{tA,tB} = 1 - \frac{1}{2} \sum_{i=1}^k |f_{tA,i} - f_{tB,i}|$$

We then calculate the measure of agreement between the cluster distributions for two visits $a_{tA,tB}$, which ranges from zero (indicating no common elements) to one (indicating identical distributions). This measure provides insight into whether the clustering algorithm captured consistent patterns in the data, particularly with regard to the patient's mental state.

In [53], the current approach has been extended. Beside SOM another algorithm has been compared called Fuzzy C-Mean. Furthermore automatic feature selection has been used, so distinct important features. More patient data has been tested in that study.

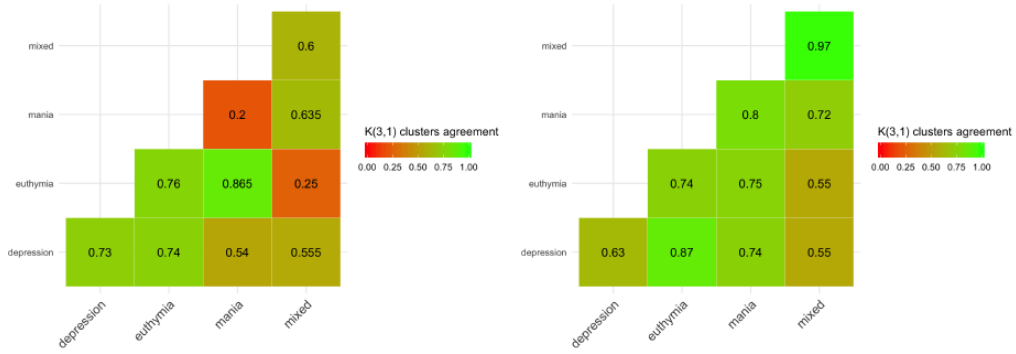


Figure 19: Degree of agreement for SOM for (left) with expert parameters and (right) automatic FS subset of acoustic parameters. [53]

Results obtained in that paper are depicted in Figure 19, where two heatmaps present the degree of agreement between clusters across four mental health states. The diagonal elements indicate agreements within the same state, aiming for values close to 1, while off-diagonal elements reflect cross-state agreement, ideally low. Left heatmap, that results were received during SOM algorithm with expert features selection submit on the diagonal agreements: Depression (0.73), Euthymia (0.76), Mania (0.2), and Mixed (0.6). While depression and euthymia show satisfactory values, mania appears insufficient due to the unique characteristics of the phase, and mixed demonstrates relatively strong performance considering its complexity. Remaining combination: Euthymia-Mania agreement (0.865) contradicts medical intuition, as the two states are fundamentally distinct. Other values, like

Mixed-Euthymia (0.25) and Mania-Depression (0.54), align better with expectations.

In the second heatmap (using automatic feature selection and SOM algorithm) we obtained diagonal agreements improved overall, with values: Depression (0.63), Euthymia (0.74), Mania (0.8), and Mixed (0.97). Notably, the mixed state achieves near-perfect agreement, while mania improves significantly from the first map. Off-diagonal agreements mostly stay above 0.5, indicating satisfactory but not ideal separation between states.

Overall, the right heatmap outperforms the left heatmap due to better diagonal agreements, particularly for challenging states like mania and mixed. However, challenges remain in minimizing off-diagonal agreements, especially in cases like Mania-Euthymia.

6.3 The Proposed Semi-supervised Validation Approach based on Cluster Agreement

The proposed method was inspired by cited previous two authors' publications [51, 53]. Previous method didn't considered some aspects like label uncertainty and treating recordings separately. Based on that assumption, a new validation method called Semi-supervised Cluster Agreement (SCVA) has been created. This approach aims to address the secondary problem of this dissertation, which is to determine the degree of similarity between labeled and unlabeled data.

Obtaining medical labels for every recording is practically impossible. The psychiatrists do not constantly monitor the patient and do not assess his or her mental state. Therefore, all data were subjected to clustering using the Self-Organizing Map (SOM) algorithm in various topological variants to identify similar audio recordings. Tested topologies are presented in Figure 20.

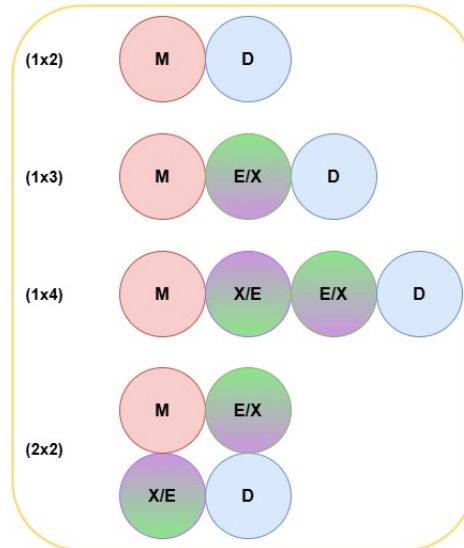


Figure 20: Self-Organizing Map topologies considered during clustering and SCVA methods

Topology (1x2) assumes that we distinguish 2 different BDs for each mobile recording - so it could be in Mania or Depression State. Topology (1x3) assumes that we can distinguish between depression and mania. The remaining values could belong to euthymia and/or mixed. The following two topologies (1x4 and 2x2) assume that each recording could belong to one of 4 available BD states- however, we still cannot indicate which indirect states are close/far from Mania and Depression.

It is important to note that **cluster values are not assigned to specific mental states** - in other words, we do not assign any interpretation, such as

Cluster 1 indicating depression. This is not crucial for us at this stage. What matters is whether the clustering itself is consistent and similar across different instances.

Clustering all available recordings may not provide clear insights into the patient's condition, so we focus only on the periods surrounding medical visits preserving labels uncertainty. An example of SOM clustering using topology 1x3 is presented in Figure 21. All of the available recordings for each patient were clustered. Then we grouped all from time frame of visit by the distance from visit date using fixed time window.

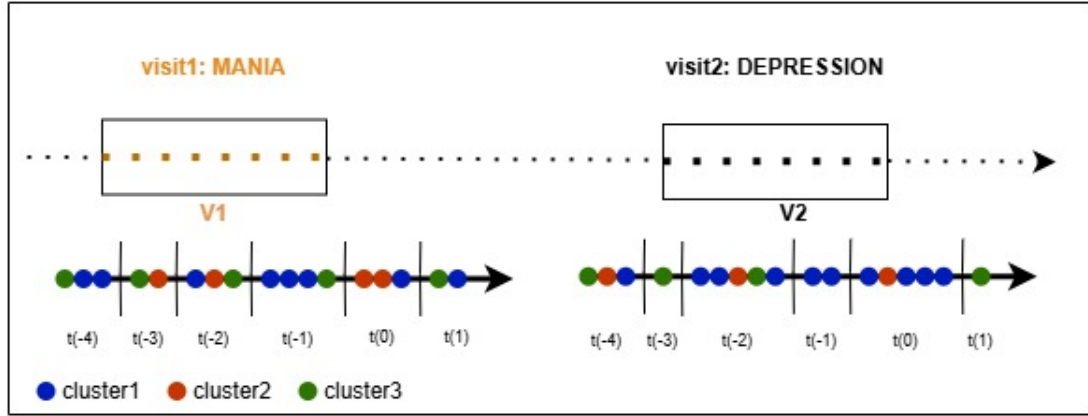


Figure 21: Clustering with Self-Organizing Maps using Topology 1x3 preserving label uncertainty

Next step is to compare the frequencies between similar cluster from selected visits with particular time window group between each of visit combination using the RAND index that is calculated using (18):

$$SCVA = \left(\frac{\sum_{k=1}^{k_{\max}} RAND(f_{1,k}, f_{2,k})}{k_{\max}} \cdot \sqrt{\frac{r_{\max}}{r_{\min}}} \right) \quad (18)$$

Where:

- f : is a vector of the frequency with received clusters k in each time step around visits (like: -7, -6,...,0,...2). The Confidence factor represented the uncertainty of label was described as in Section 2.3
- r : is a ratio between the number of mobile recordings during the extended window of both visits
- $RAND$: is defined by $RAND = \frac{a+b}{\binom{n}{2}}$

An Example of calculated frequencies f from the example presented in Figure 21 is presented in Table 5. Column Confidence Factor and Time Point are related to the distance from the visit. The next three columns show the frequencies of particular clusters in the first visit, and the last three columns show the frequencies of clusters from the second visit. Based on that example, the final value of SCVA is equal to 0.773.

The agreement between clusters is crucial for assessing the predictive potential of clustering. This involves determining whether the clusters are associated with the phases of bipolar disorder rather than other irrelevant features of the patient's speech.

Table 5: Frequencies of clusters presented in Figure 21

Confidence Factor	Time Point	Visit 1, Frequencies of:			Visit 2, Frequencies of:		
		Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
1	0	1	2	0	4	1	0
0.9	-1, +1	4	0	2	2	0	1
0.8	-2, +2	1	1	1	3	1	1
0.7	-3	0	1	1	0	0	1
0.6	-4	2	0	1	1	1	1
0.5	-5	0	0	0	0	0	0
0.4	-6	0	0	0	0	0	0
0.3	-7	0	0	0	0	0	0

The obtained coefficients are applied to a heatmap (see Figure 22), averaging the results for similar mental states. This matrix indicates the similarity of the clusters within similar visits, showing whether the clustering algorithm assessed the recordings similarly when the patient was in a comparable mental state.

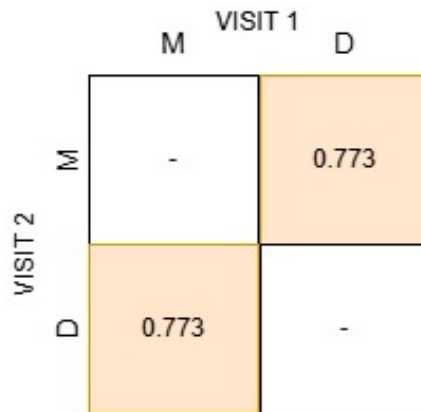


Figure 22: SCVA heatmap

Several factors can influence this process. Most patients did not experience all phases of BD during the study period, which means the clustering algorithm may not extract symptoms from unobserved phases (e.g., depression) and may instead create clusters within the observed data, such as splitting manic states into hypomania and full-blown mania. Additionally, the heterogeneity of symptoms in a mixed state may cause data from this phase to be distributed across different clusters. Outliers can also disrupt unsupervised learning methods, mainly when a predefined number of clusters is used. Moreover, the arrangement of clusters within a 3x1 SOM map can vary, leading to different cluster orders (e.g., depression x euthymia x mania or mania x euthymia x depression).

Finally, further questions arise about whether clustering should be performed separately for each patient or collectively for all patients. The study involved patients with bipolar disorder, but each patient has unique characteristics during specific periods of the disease, which may differ significantly from other patients, potentially destabilizing the model. Additionally, a comparison by gender could be considered to generalize the model better. There are many factors to consider, but it is crucial to balance the available data with the study's objectives.

7 Results

7.1 BDmon system and data transformation

Our research originates from the project entitled “Smartphone-based diagnostics of phase changes in the course of bipolar disorder.” In that project, around 70 patients diagnosed with bipolar disorder were involved. They received a dedicated mobile application called BDMon, which collected features such as the number of incoming and outgoing calls, the length of SMS messages, and the number of steps. Furthermore, it implemented a version of the OpenSmile [23] library, which collects the physical features of voice. Each patient was associated with a psychiatrist and met with them at frequent appointments. Information about the patient’s current mood was obtained using the Hamilton Rating Scale for Depression and the Young Mania Rating Scale to assess the two dimensions of the disease.

The sound signal is obtained directly from the microphone to avoid recording the interlocutor’s speech. The voice signal is processed in real-time to extract its physical descriptors, which are transferred to a secure server and stored in a database for analysis. The cached data are permanently deleted from the phone after feature extraction. Therefore, sensitive data do not leave the patient’s phone, and the voice parameters sent to the server do not allow for identifying the patient’s identity.

Data transformation is a crucial step in the processing and analysis of acoustic data, where raw audio signals are converted into a format that is more suitable for analysis, visualization, or further processing. This process often involves several key techniques and methodologies designed to enhance the signal-to-noise ratio, extract meaningful features, and prepare the data for applications such as machine learning, sound classification, and acoustic pattern recognition.

In this study, we work with acoustic data obtained from recordings of patients collected in their neutral environments. The data comprises 10–20 ms frames that capture 86 acoustic parameters. We consider two approaches to data preparation in this dissertation:

Raw Data Approach: The first approach (presented in Section 1.3 as *A1*) involves using the raw data for each conversation. That dataset was prepared separately for this dissertation. Patients were selected who had multiple visits and, in particular, exhibited different states and symptoms of the disease during those visits. Specifically, we extract the first 1,000 frames from each conversation, describing each frame by 86 acoustic parameters. This method preserves the detailed information present in the raw acoustic signals, allowing for a granular analysis of the vocal characteristics. The number of frames was selected to balance the informativeness of recording and computational complexity during calculation.

Aggregated Data Approach: The second approach (presented in Section

1.3 as A_2) aggregates the data from each conversation by calculating the mean value for all acoustic parameters. That dataset was developed during the BIPO-LAR project, where one of the artifacts was precisely such a dataset. This results in a single averaged set of parameters representing each conversation. In another of my works, various data aggregation methods were explored, and the mean was recognized as the leading method for summarizing acoustic data [49]. This approach simplifies the dataset, reducing computational demands and emphasizing the overall trends within each conversation.

By employing these two approaches, we aim to assess the impact of data granularity on the analysis and interpretation of acoustic features in the context of bipolar disorder.

7.2 Hardware and Software

All of the experiments were prepared using Notebook DELL Precision 7710 with System Operation Windows 10.

CPU: Intel Core i7 - 6920HQ CPU 2.9GHz

GPU: NVIDIA Quadro P3000

RAM: 64GB

For software development, used VS Code as a primary platform for Python 3.10, and RStudio for R language in version 4.2

All of codes are available in the GITHUB platform:

PYTHON: https://github.com/kaminskaolga/phd_NN

R: https://github.com/kaminskaolga/chad_prep

Moreover, some preprocessing components presented in this work were additionally implemented in the **bipolar** package aimed for processing sensor and psychiatric data. The open repository is available under the following link on GitHub: <https://github.com/ITPsychiatry/bipolar>. To my current knowledge, there is no programming package that supports processing sensor and psychiatric data. The package serves as the modeling engine used in uncertainty-aware models. The software offers two main features: modeling and visualization. Users have a range of functions enabling such tasks as data preparation, configuration, and running experiments with different parameters.

The other part is a dashboard where users can preview their datasets and model results. They can easily review performance metrics of a model, display history of patients' visits etc. That software is delivered in form of an *R* package.

The idea behind the *bipolar* package is presented in Figure 23. The process starts with the usage of two data sources: **sensors** - unlabeled sensor data and **visits** - ground-truth labels from psychiatric diagnoses.

Next, preprocessing is applied. This involves cleaning the raw sensor data and

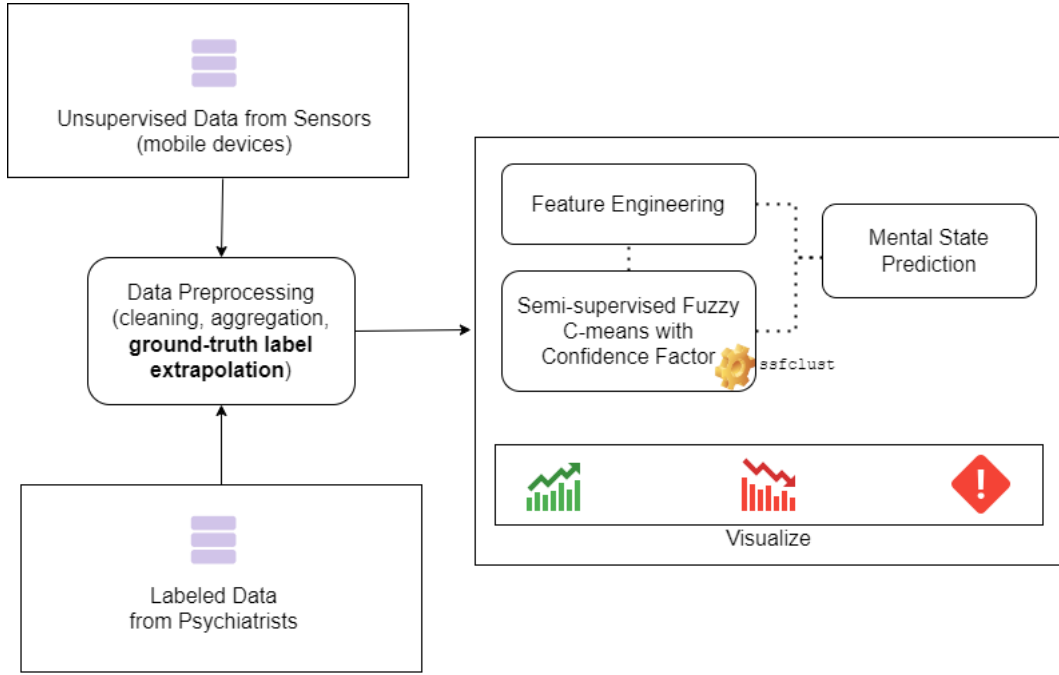


Figure 23: The data flow in *bipolar* package

its Aggregation to reduce the frequency of the data and make it more manageable over specific time intervals. Psychiatric labels, which are often collected less frequently, are aligned with sensor data through ground-truth extrapolation, assuming the labels persist over time.

Following preprocessing, feature engineering begins that consist of feature selection that is used to identify the most relevant features, reducing data complexity and improving model performance.

Once the features are prepared, semi-supervised learning is applied. This step involves training models on both labeled data (psychiatric assessments) and a more extensive set of unlabeled data (sensor readings). Semi-supervised learning algorithms learn to predict the patient's mental state based on the patterns found in the sensor data and psychiatric labels. Finally, visualization techniques are used to display the results and manage them. Figure 24 presents an example of that visualization.

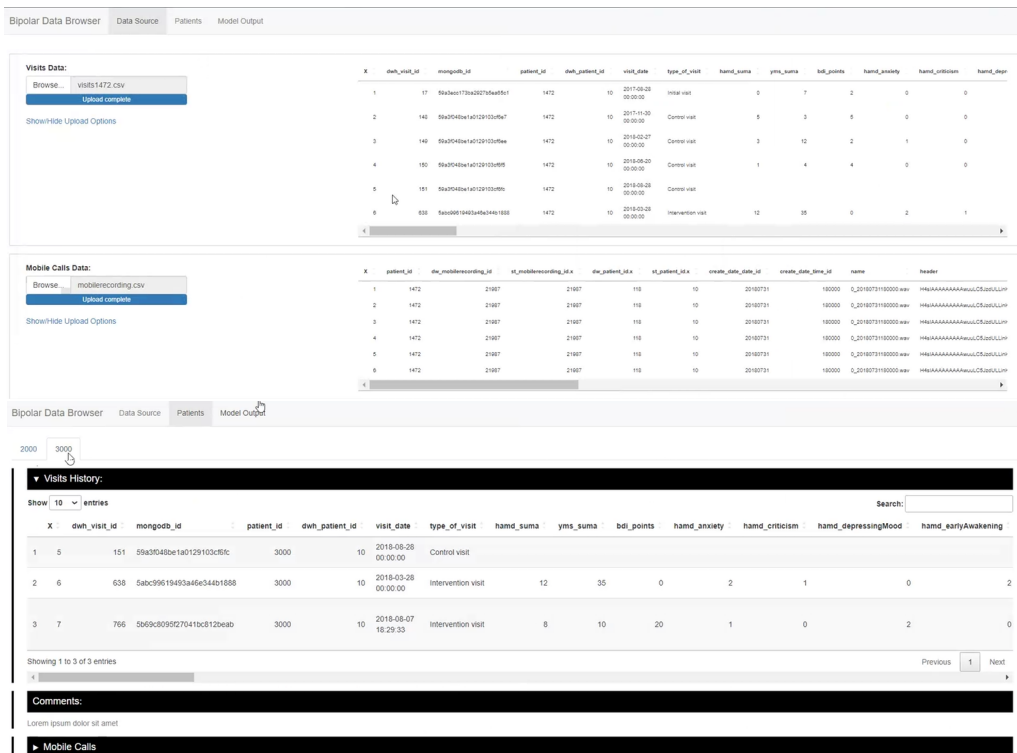


Figure 24: Example of dashboards received by *bipolar* package.

7.3 Numerical results

I now present the numerical results for the major novel approaches proposed in this dissertation. In subsection 7.3.1 and 7.3.2 I will present methods became from Section 4 about Feature Selection, more precisely Uncertainty Aware Advanced Feature Selection (UAAFS) and Cost-Constrained Feature Selection (CCFS). Next subsection 7.3.3 and 7.3.4 present results described in Section 5 about the implementation of Neural Networks and method called Explainable Compositional Multilayer Perceptron with Confidence Factor (XCMLPCF). In subsection 7.3.5 present results from Section 6 with Semi-supervised Cluster Validation Agreement (SCVA) method. Finally, in subsection 7.4, there is a discussion about received results.

7.3.1 Uncertainty Aware Advanced Feature Selection Method

A method called Uncertainty Aware Advanced Feature Selection (**UAAFS**) reflects the problem of feature selection of that dissertation. It is based on vector W_{rp} , where data were aggregated into each mobile recording. The labels are presented as a vector within label confidence. We considered three types of Label Confidence Scenarios. First one assumes that labels around visits accurately 7 days before visit, the day of the visit - 'day 0' and 2 days after visits are equally important - equally weighted and its called *constant*. In second and third scenario we considered labels uncertainty and it could be modeled as were mentioned in Section 2.3 respectively *gauss* and *step* approach. For each of label scenario we considered three types of feature selection approach. In first we are taking into account all available data therefore all 4 gathered BD states (euthymia, depression, mania, mixed). In second and third we distinguish important features related to type of Bipolar Disorder States precisely Depression symptoms (selecting depression and euthymia) and Maniac symptoms (selecting maniac and euthymia). That kind of approach is important to compare and explore the most important features of deviated BD states. All scenarios are presented in Table 6

Table 6: The list of scenarios that were created to verify the UAAFS method

No.	Label Confidence	Type
1	gauss	4 class: euthymia vs depression vs mania vs mixed
2		2 class: depression vs euthymia
3		2 class: mania vs euthymia
4	step	4 class: euthymia vs depression vs mania vs mixed
5		2 class: depression vs euthymia
6		2 class: mania vs euthymia
7	constant	4 class: euthymia vs depression vs mania vs mixed
8		2 class: depression vs euthymia
9		2 class: mania vs euthymia

The dataset and UAAFS feature selection method for the presented algorithm were prepared using *bipolar* package (mentioned in Section 7.2). The first step is preparing a dataset that contain label confidence for 3 scenarios. Next, we processed the UAAFS method that returns the 10 most important features for each scenario. The obtained results are present in Table 7 to distinguish the best features representing depressive symptoms and in Table 8 that represents maniac symptoms.

Important parameters in Depression presented in Table 7 has a high degree of similarity, with several features appearing in all three categories. Key common features include `audSpec_Rfilt_sma_compare_5_`, `audSpec_Rfilt_sma_compare_22_`, `flfrequency_sma3nz`, `pcm_fftMag_mfcc_5_`, `pcm_fftMag_mfcc_9_`, and `pcm_fftmag_spectralminpos_sma`. Additionally, some features are shared between two columns, such as `audSpec_Rfilt_sma_compare_21_` and `slope0.500_sma3` in Gauss and Step, as well as `pcm_fftMag_mfcc_7_` in Step and Constant. However, there are also unique features specific to each category.

Table 7: List of important features for Depression with considered label confidence function

	Gauss	Step	Constant
1	audSpec_Rfilt_sma_compare_22_	audSpec_Rfilt_sma_compare_5_	audSpec_Rfilt_sma_compare_5_
2	f1frequency_sma3nz	pcm_fftmag_spectralminpos_sma	pcm_fftmag_spectralminpos_sma
3	pcm_fftMag_mfcc_5_	pcm_fftMag_mfcc_5_	pcm_fftMag_mfcc_5_
4	audSpec_Rfilt_sma_compare_5_	audSpec_Rfilt_sma_compare_22_	f1frequency_sma3nz
5	audSpec_Rfilt_sma_compare_21_	f1frequency_sma3nz	pcm_fftMag_mfcc_9_
6	slope0.500_sma3	pcm_fftMag_mfcc_9_	pcm_fftMag_mfcc_7_
7	audSpec_Rfilt_sma_compare_25_	audSpec_Rfilt_sma_compare_21_	logRelF0.H1.H2_sma3nz
8	pcm_fftMag_mfcc_9_	slope0.500_sma3	audSpec_Rfilt_sma_compare_22_
9	f2frequency_sma3nz	pcm_fftMag_mfcc_7_	f0_sma
10	pcm_fftmag_spectralminpos_sma	pcm_fftMag_mfcc_8_	audSpec_Rfilt_sma_compare_4_

Important parameters in Mania presented in Table 8 share several key features, including f0_sma, pcm_fftMag_mfcc_4_, pcm_fftMag_mfcc_5_, logRelF0.H1.H2_sma3nz, f1bandwidth_sma3nz, pcm_fftMag_mfcc_7_, and pcm_fftMag_mfcc_9_. Additionally, the feature pcm_fftMag_fband0.250_sma appears in all categories, although the slope features differ slightly: Gauss includes slope0.500_sma3, while both step and Constant feature slope500.1500_sma3. There are a few features that are present in two out of the three categories. For instance, pcm_fftMag_mfcc_8_ is found in both step and Constant, but not in Gauss. Conversely, Gauss uniquely includes pcm_fftMag_mfcc_2_, which does not appear in the other categories.

Table 8: List of important features for Mania with considered label confidence function

	Gauss	Step	Constant
1	f0_sma	f0_sma	f0_sma
2	pcm_fftMag_mfcc_4_	pcm_fftMag_mfcc_4_	pcm_fftMag_mfcc_7_
3	pcm_fftMag_mfcc_5_	pcm_fftMag_mfcc_7_	logRelF0.H1.H2_sma3nz
4	pcm_fftMag_mfcc_2_	pcm_fftMag_mfcc_5_	pcm_fftMag_mfcc_5_
5	f1bandwidth_sma3nz	logRelF0.H1.H2_sma3nz	pcm_fftMag_mfcc_4_
6	logRelF0.H1.H2_sma3nz	f1bandwidth_sma3nz	pcm_fftMag_mfcc_8_
7	pcm_fftMag_mfcc_7_	pcm_fftMag_mfcc_9_	f1bandwidth_sma3nz
8	pcm_fftMag_mfcc_9_	slope500.1500_sma3	pcm_fftMag_mfcc_9_
9	pcm_fftMag_fband0.250_sma	pcm_fftMag_mfcc_8_	slope500.1500_sma3
10	slope0.500_sma3	pcm_fftMag_fband0.250_sma	pcm_fftMag_fband0.250_sma

Having important features in Bipolar Disease we need to verify the power of classification using that features set. For that study, we focus on well-known classifiers: XGBoost (described in Section 4.2.3), SVN and KNN (both described in Section 3.2.1). The classification process were conducted using previous dataset with selected features for each mentioned scenario. For the training set data except the day of the visit were selected, and for test set the data coming from the day of visit (as the most reliable). The proportion of train and test data are respectively 88% to 12% - details statistics are presented in Table 9.

Table 9: Train and test sets data distribution

	Euthymia	Depression	Maniac	Mixed	Σ
Train	1471	1394	748	3884	7497
Test	227	195	68	46	536
Σ	1698	1589	816	3930	8033

The results of each type of classification are presented in Tables 10, 11, 12.

In the four-class classification task aimed at distinguishing between depression, mania, euthymia, and mixed, the performance of XGBoost, SVM, and KNN models was evaluated across three labels confidence approach: Gauss, Step, and Constant. Results are presented in Table 10.

Table 10: Classification results for 4 classes using XGB, SVM, and KNN

4 class	Accuracy			F1			Precision			Recall		
Type	XGB	SVM	KNN	XGB	SVM	KNN	XGB	SVM	KNN	XGB	SVM	KNN
Gauss	0,56	0,59	0,49	0,61	0,67	0,55	0,61	0,58	0,52	0,61	0,78	0,57
Step	0,60	0,58	0,48	0,64	0,66	0,55	0,65	0,58	0,52	0,63	0,75	0,57
Constant	0,63	0,60	0,45	0,66	0,67	0,51	0,66	0,58	0,48	0,67	0,79	0,55

- **The Constant** approach was the best overall performer, particularly benefiting XGBoost. It achieved the highest accuracy (0.6306), F1 score (0.6633), and precision (0.6550) among all distributions. SVM also performed well under the Constant distribution, with accuracy (0.6007) and the highest recall (0.7885). This suggests that Constant distribution provides a balanced and stable performance for multi-class problems, especially for XGBoost and SVM.
- **The Gauss** approach offered solid results in terms of recall, with SVM achieving the highest recall of all distributions (0.7797) and XGBoost showing strong recall as well (0.6051). However, the accuracy and F1 scores were lower compared to Constant and Step, indicating that while Gauss is suitable

for capturing positive cases (high recall), it might struggle with balancing other metrics.

- **The Step** approach performed reasonably well but did not outperform Constant. XGBoost achieved moderate accuracy (0.5970) and F1 score (0.6390), while SVM had slightly lower results compared to its performance in Gauss and Constant. KNN lagged behind in all metrics across the distributions, with its best recall (0.5727) coming from the Step distribution but consistently lower performance elsewhere.

In the binary classification task for depression vs euthymia, the performance of XGBoost, SVM, and KNN was evaluated under Gauss, Step, and Constant labels confidence approach across four metrics: accuracy, F1 score, precision, and recall. Results are presented in Table 11.

Table 11: Classification results for 2 (euthymia vs depression) class using XGB, SVM, and KNN

DEPRESSION	Accuracy			F1			Precision			Recall		
Type	XGB	SVM	KNN	XGB	SVM	KNN	XGB	SVM	KNN	XGB	SVM	KNN
Gauss	0,77	0,76	0,68	0,74	0,78	0,71	0,76	0,76	0,70	0,73	0,81	0,73
Step	0,72	0,78	0,67	0,70	0,80	0,69	0,70	0,78	0,70	0,70	0,82	0,68
Constant	0,73	0,77	0,66	0,71	0,79	0,67	0,72	0,79	0,69	0,70	0,79	0,65

- Gauss distribution yielded strong overall performance, particularly for SVM, which achieved the highest recall (0.8062) and F1 score (0.7804). XGBoost also performed well under Gauss, with the highest precision (0.7566) and good recall (0.7333). KNN, however, lagged behind, with its highest performance metrics still being relatively low compared to XGBoost and SVM.
- Step distribution gave the best results for SVM, with the highest accuracy (0.7796), F1 score (0.7991), and recall (0.8150) of all models. XGBoost and KNN saw slightly weaker performance under Step compared to Gauss, with XGBoost achieving lower accuracy (0.7227) and F1 score (0.6992). KNN had the lowest performance in all metrics under this distribution.
- Constant distribution provided moderate performance, with XGBoost showing improvement in accuracy (0.7322) and F1 score (0.7065) over Step. SVM performed similarly to its performance in Gauss and Step, though its recall dropped slightly (0.7930). KNN had its lowest performance overall in this distribution.

SVM excelled across all distributions, with its best results under the Step distribution. XGBoost performed well, especially in Gauss, where it achieved high precision and recall. KNN consistently underperformed across all distributions, with its highest scores under Gauss.

For the binary classification of mania vs euthymia, the performance of XGBoost, SVM, and KNN was again compared across the Gauss, Step, and Constant labels confidence approach. Results are presented in Table 12.

Table 12: Classification results for 2 classes (euthymia vs mania) using XGB, SVM, and KNN

MANIA	Accuracy			F1			Precision			Recall		
Type	XGB	SVM	KNN	XGB	SVM	KNN	XGB	SVM	KNN	XGB	SVM	KNN
Gauss	0,79	0,77	0,72	0,86	0,85	0,82	0,86	0,84	0,80	0,87	0,86	0,85
Step	0,83	0,77	0,74	0,89	0,85	0,84	0,88	0,84	0,81	0,89	0,86	0,87
Constant	0,81	0,77	0,73	0,88	0,85	0,83	0,86	0,84	0,80	0,90	0,86	0,86

- Step distribution provided the strongest performance for all three models, with XGBoost showing the highest accuracy (0.8271), F1 score (0.8884), precision (0.8826), and recall (0.8943). SVM also performed well under this distribution, with an F1 score of 0.8515 and a precision of 0.8442. KNN showed its best performance here as well, with its highest accuracy (0.7424) and recall (0.8678).
- Gauss distribution showed slightly weaker performance than Step but still yielded high recall for SVM (0.8590) and XGBoost (0.8722). KNN performed moderately well under Gauss, achieving its second-best scores for accuracy (0.7220) and F1 score (0.8248).
- Constant distribution delivered competitive performance for XGBoost and SVM, with XGBoost achieving strong recall (0.8987) and F1 score (0.8793). SVM performed similarly across all three distributions, with minor differences in recall and precision. KNN performed moderately in this distribution, with the highest recall (0.8634) but generally lower performance compared to Step.

Step distribution was the best overall, producing the highest scores for all models, particularly XGBoost, which excelled in all metrics. Gauss distribution was close behind, with SVM performing well, especially in terms of recall. Constant distribution was also strong, particularly for XGBoost in recall, though it slightly lagged behind Step for overall performance.

Across the multi-class classification task (depression, mania, euthymia, and mixed) and the binary tasks (when we distinguish 2 scenarios: depression vs euthymia and mania vs euthymia), the Step and Gauss distributions tended to provide strong, consistent results. However, for the multi-class task, the Constant distribution emerged as the best performer overall. It provided the highest accuracy, F1 score, and precision, especially benefiting XGBoost and SVM.

Constant distribution was particularly strong in precision and F1 score for XGBoost, making it the top choice for balanced performance in the multi-class classification task. Gauss distribution effectively boosts recall, especially for SVM, though its overall performance in accuracy and F1 score was not as competitive. Step distribution showed promising results but was less effective than Constant in the multi-class task, although it did provide solid recall for SVM and KNN.

In binary classification tasks (when we distinguish 2 scenarios: depression vs euthymia and mania vs euthymia), Step and Gauss distributions typically outperformed Constant in terms of accuracy, F1 score, and precision, making them the better options for balanced performance. However, in the more complex four-class classification task, the Constant distribution proved to be the best overall. XGBoost and SVM achieved their highest scores under the Constant distribution, particularly excelling in precision and F1 score.

Thus, while Gauss and Step distributions are generally better for binary tasks and provide robust performance, the Constant distribution can outperform them in multi-class settings where overall balance across metrics like accuracy, precision, and recall is critical.

7.3.2 Cost-Constrained Feature Selection

The primary objective of the experiments was to evaluate the cost-constrained feature selection algorithm against its conventional version. That problem is still associated with problem of feature selection of that dissertation. First, we individually optimized the parameter λ for each budget. Next, we executed both the traditional and the cost-constrained feature selection algorithms using $\lambda = 0$ and $\lambda = \lambda_{\text{opt}}$, respectively. Finally, we assessed the performance of a Random Forest [2] model, which was trained using the selected features, by computing its AUC score. All optimization, feature selection, and model training were carried out on the training set (50% of the data), while the AUC metric was computed using the test set (the remaining 50%).

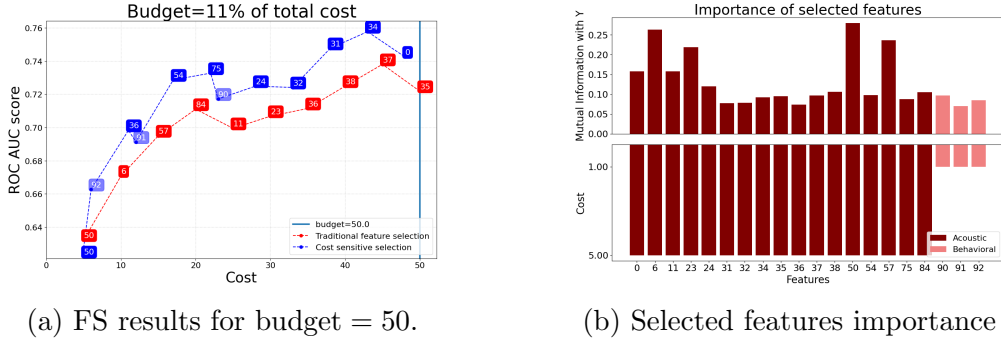


Figure 25: Cost Constrained feature selection results

Figure 25a presents the AUC of 2 models (blue line using CCFS, and red line using traditional feature selection methods) and in Figure 25b we can see the importance of these features measured by mutual information with the target variable.

Acoustic features exhibit higher mutual information but are costly, whereas behavioral features are five times cheaper with lower mutual information. In both algorithms' initial step, feature number 50 (magnitude of fast Fourier transform coefficients in the 250-650Hz band) from the acoustic group is selected due to its high informativeness. In the second step, the cost-constrained method selects feature number 92 (ratio of outgoing calls to all mobile calls per day) from the behavioral group, while the traditional method selects feature 6 (another fast Fourier transform coefficient magnitude) from the acoustic group. In the third step, both methods choose acoustic features. As the budget increases, both methods incorporate more features, but the cost-constrained method consistently yields a higher AUC score, demonstrating superior performance in classifying BD episodes.

The preliminary results indicate that applying the cost-constrained method enhances the accuracy of the classification model in the context of budget-constrained

BD episode classification problems. This improvement in predictive performance is particularly pronounced under low-budget conditions. In contrast, as the budget increases, the predictive power of both traditional and cost-constrained methods tends to converge.

7.3.3 Primary Neural Network Classification

The research focused on comparing different architectures of the following selected popular Neural Networks: ANN, CNN and LSTM. For that task we divided all labeled data into 3 datasets (as visualized in Figure 26). For training we are using 89% of data from surrounding days of visit. Additionally this subset distinguishes the validation set (22,3%) to estimate how well your model has been trained. The rest 10,7% of it is test set that is used to validate NN. All results are tested on day of the visits as the most certain data with received label. Data distribution of 3 datasets are presented in Table 13.

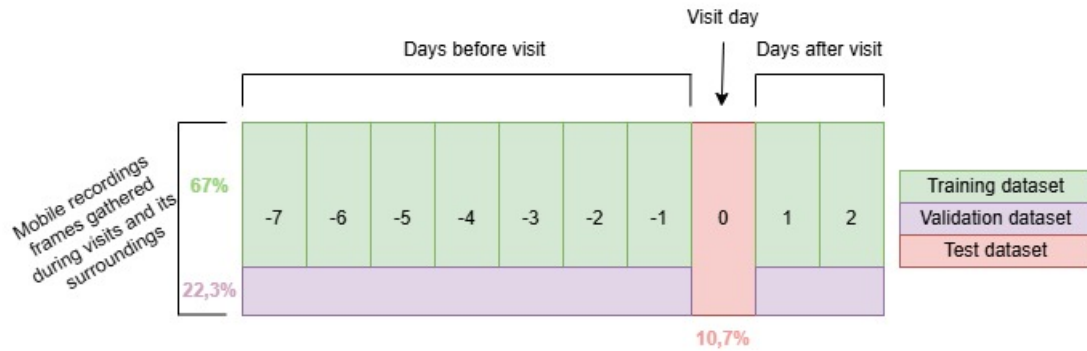


Figure 26: Division of collected data into training, validation and test subsets

Table 13: Train, Valid and Test Data distribution

	Euthymia	Depression	Maniac	Mixed	Σ
Train	536	469	494	284	1783
Valid	179	156	165	95	595
Test	77	104	63	42	286
Σ	792	729	722	421	2664

Introduced in Section 5, three architectures (ANN, CNN, LSTM) were trained to classify BD state.

In Figure 27 the line plot compares the loss trends for each model over training epochs, with both training and validation losses represented: CNN model shows a relatively higher initial loss but steadily decreases over epochs. It converges with some fluctuations in validation loss after the 10th epoch. LSTM exhibits a lower starting loss than CNN but converges more slowly. It has a more consistent validation trend compared to CNN. ANN starts with a much lower loss and remains relatively stable over epochs, indicating a simpler model that quickly reaches its optimal performance plateau.

Final performances were presented in Table 14. An ANN model achieved a moderate accuracy (0.50) with the lowest loss (1.16) among the models, showing balanced metrics but lower performance compared to CNN. CNN has the highest accuracy (0.60) but also the highest loss (1.45), along with the best F1-score (0.61). This indicates it generalizes slightly better despite higher training loss. LSTM oscillates between ANN and CNN with accuracy at 0.48 and a loss of 1.35, achieving intermediate values for precision, recall, and F1-score. It could capture sequential patterns better but does not outperform CNN. The CNN appears to perform best overall for this task, achieving the highest accuracy and F1-score, making it more reliable for classification.

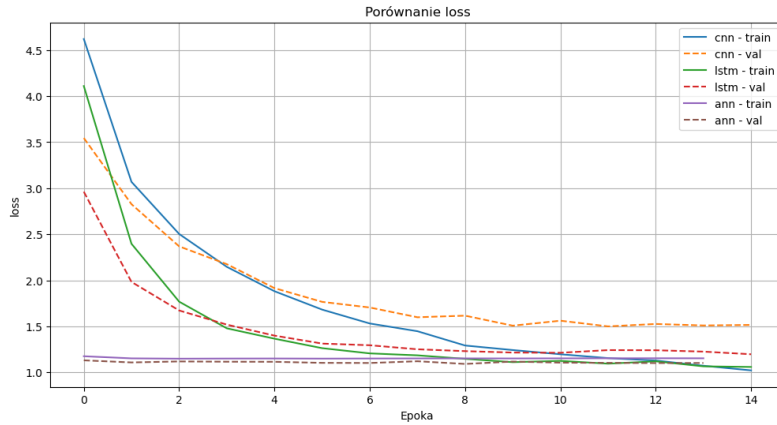


Figure 27: LOSS for 3 architecture

Table 14: Detailed results for 3 different type of NN architecture. Details information like layers or hyperparameters are available in Section 5.1

Model	Accuracy	Loss	Precision	Recall	F1-score	time[s]
ANN	0.50	1.16	0.53	0.50	0.49	5244s
CNN	0.60	1.45	0.62	0.60	0.61	59s
LSTM	0.48	1.35	0.48	0.48	0.47	103s

Therefore we present the ROC AUC for each class for the CNN model and its presented in Figure 28. The CNN model shows the highest discriminative power for Depression (AUC = 0.88), where it performs best in distinguishing it from other BD states. Mania and Mixed also have good AUC scores (0.79 and 0.80, respectively), indicating acceptable performance. Euthymia has the lowest AUC score (0.74), suggesting that the model struggles the most with distinguishing this class, potentially due to overlapping features with other classes. The diagonal line represents a random classification (AUC = 0.50), and all classes are above this baseline, confirming that the CNN model is effective to varying degrees across classes.

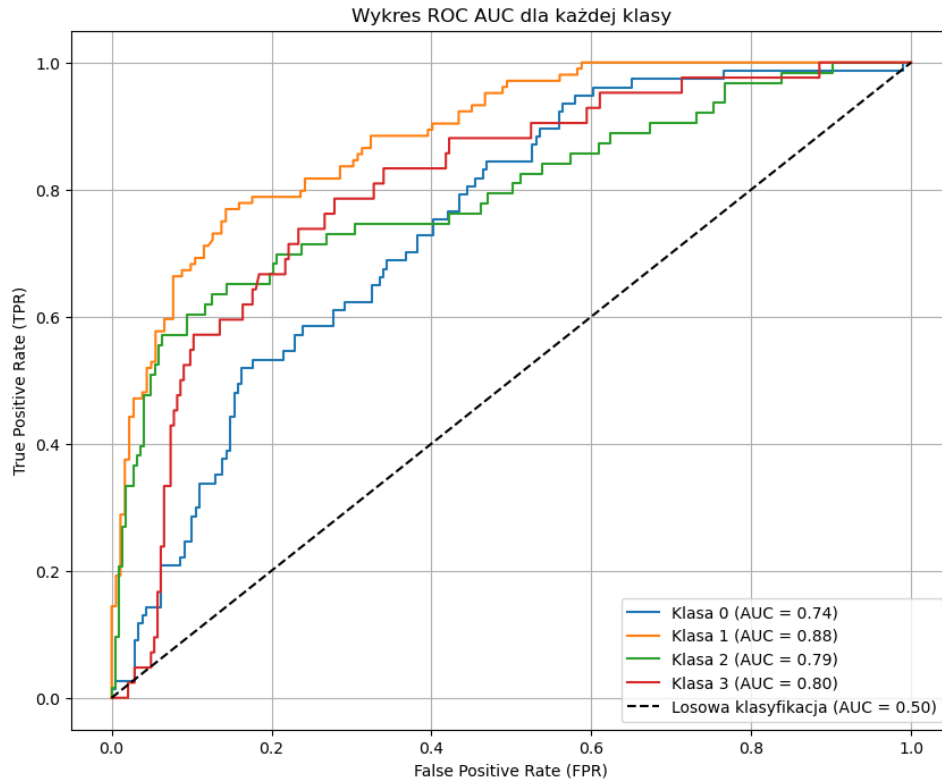


Figure 28: The AUC ROC for the CNN. Class 0 represents euthymia, 1 - depression, 2 - mania, 3 - the mixed state.

7.3.4 Explainable Compositional Multilayer Perceptron

The goal of experiments performed in this Section is two-fold:

1. to validate the proposed confidence-based neural network approach
2. to investigate experimentally the accuracy and explainability of the proposed two-layer output architecture including not only bipolar disorder classes (euthymia, depression, mania, mixed states), but also the considered symptoms as presented in Table 2.

The major method called Explainable Compositional Multilayer Perceptron **XCMLP** presented in Section 5.2 were performed in 2 ways:

- as a basic XCMLP method
- as a Explainable Compositional Multilayer Perceptron with Confidence Factor (**XCMLPCF**) considering *STEP Label Confidence*.

The Neural Network architecture is trained using dataset with train, valid and test dataset similar as was mentioned in Figure 26.

Table 15 presents detailed metrics for that NN. The XCMLP model without weights achieves better metrics overall, with an accuracy of 0.51, a precision of 0.54, recall of 0.46, and an F1-score of 0.45, along with a lower loss of 0.9. In contrast, the XCMLPCF model trained with the Step confidence factor has weaker performance, with an accuracy of 0.46, precision of 0.47, recall of 0.46, and an F1-score of 0.45, and a higher loss value of 1.27.

Table 15: Detailed results received with PLENARY architecture

Model	Accuracy	Loss	Precision	Recall	F1-score	time[s]
XCMLP	0.51	0.90	0.54	0.46	0.45	6958s
XCMLPCF with step confidence factor	0.46	1.27	0.47	0.46	0.45	7716s
Hyperparameters: optimizer: ADAM, learning_rate = 0.002, epochs : 15, batch_size : 32						

However, the model trained with weights shows better overall performance in terms of loss reduction and stability during training and validation - see Figure 29. Its training loss curve starts lower and decreases steadily, while the validation loss remains relatively stable throughout the epochs, although slightly higher than the training loss.

The next step is the evaluation of ROC-AUC plots that provide insight into the discrimination ability of both models for each class. Figure 30 present the diagram for XCMLP (left) and XCMLPCF(right). For the left model (XCMLP), the highest

AUC scores were achieved for class Depression and Mania indicating the strongest performance in distinguishing these two classes compared to other classes. For the right model (XCMLPCF), the AUC scores are slightly lower. The performance drops across most classes, particularly for "Euthymia" and "Depression," where the AUC decreases noticeably.

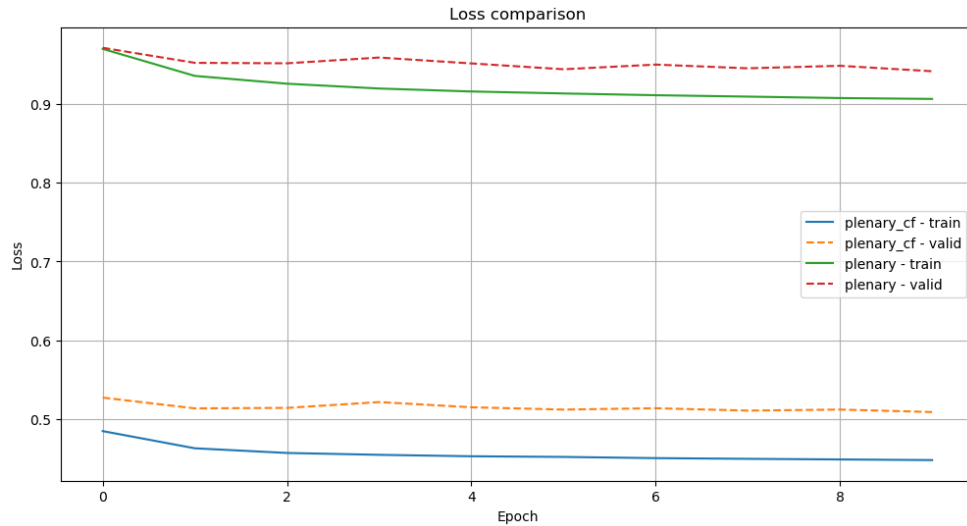
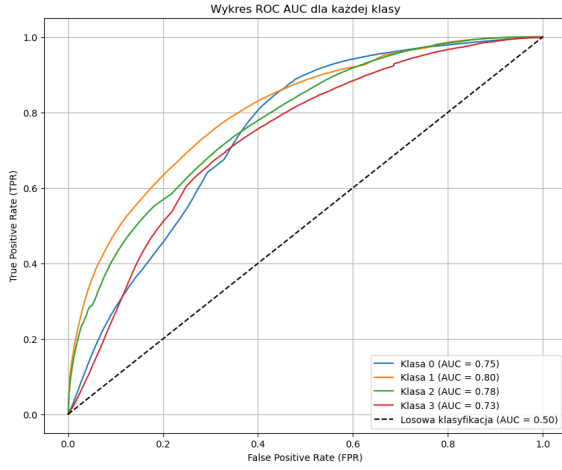
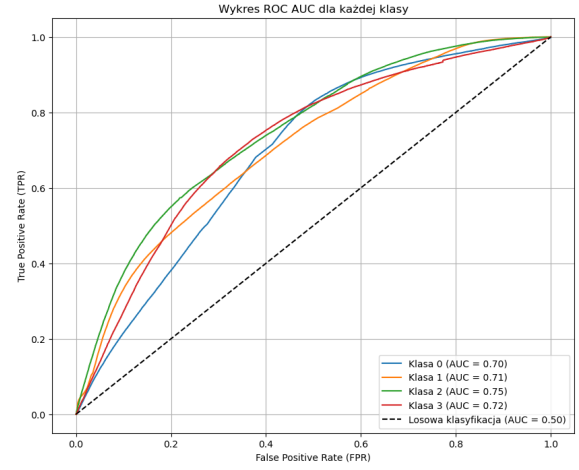


Figure 29: The loss function plot for the XCMLP model (green and red) and XCMLPCF model (blue and orange)

The most interesting results were noticed in Confusion Matrix presented in Figure: 31. The unweighted model (XCMLP) achieves better classification for the Depression, correctly identifying 70619 samples and Mania 19154 samples. The remaining classes Euthymia and Mixed achieved a better fit with the XCMLPCF model. Euthymia received 46031 correctly classified values and mixed were correctly classified 10088 times. Total Accuracy compared all classes therefore model XCMLP received better results. When looking at that diagram misclassify of Euthymia instead of Depression is noticeable. That could be caused by uneven label (more precisely, time points) distribution.

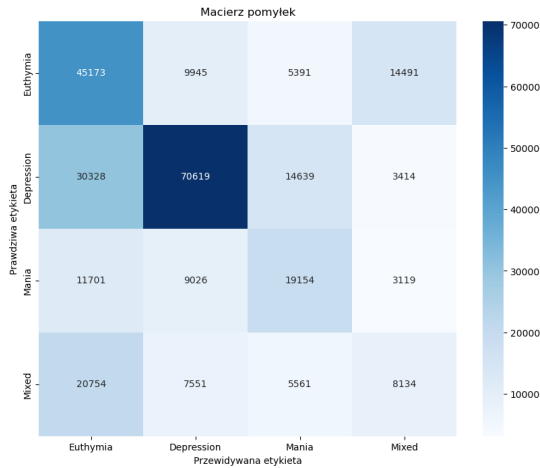


(a) Standard XCMLP method without label confidence factor

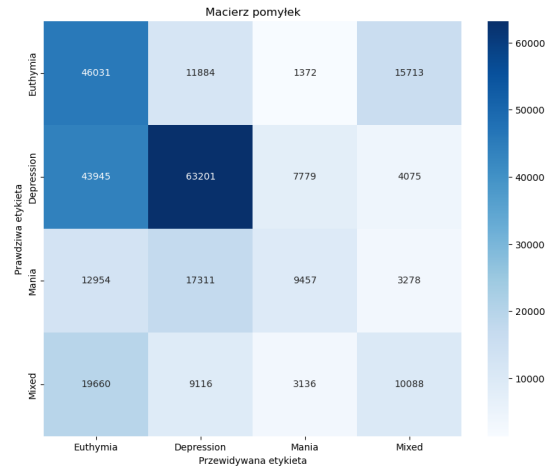


(b) The XCMLPCF method using label confidence factor

Figure 30: The ROC AUC curve for each of classes where 0 is euthymia, 1 is depression, 2 is mania and 3 is mix.



(a) Standard XCMLP method without label confidence factor



(b) The XCMLPCF method using label confidence factor

Figure 31: Confusion Matrix for those two considered methods.

The last step in XCMLP model is explainability of model results. Figure [32, 33, 34, 35] show the global explanation for BD states classification and Figure responds for symptoms. The color bars represent the average impact of the 15 most important features on the magnitude of the model output, received with the NN.

According to Figure 32 The feature `pcm_fftMag_mfcc_12_` stands out as the most influential, with the highest mean SHAP value, indicating its significant contribution to the model's decision-making across all classes. Other prominent features include `pcm_LOGenenergy_sma`, `pcm_zcr_sma`, and `f0env_sma`, which also play crucial roles in distinguishing between the states.

For instance, feature `pcm_fftMag_mfcc_12_` heavily influences predictions for both depression and euthymia, while `pcm_zcr_sma` demonstrates varied importance across all states, with a notable contribution to mania and mixed states. Additionally, features like `pcm_fftMag_psysharpness_sma_compare` and `pcm_fftMag_spectralRollOff75_0_sma` have more balanced contributions across all classes, indicating their more generalized influence.

The SHAP summary plots (depicted in Figure 33) for depression and mania reveal distinct patterns in how features contribute to the classification of these states, with some overlaps in important features but notable differences in their magnitude and direction of impact.

For depression, `pcm_fftMag_mfcc_12_` is the most impactful feature, with high values (red) strongly associated with predictions of depression. Similarly, `pcm_zcr_sma` also plays a critical role, but its effect is more balanced, with both high and low values contributing positively or negatively to the prediction.

In contrast, for mania, `pcm_LOGenenergy_sma` emerges as the most influential feature, with high values contributing strongly to the prediction of mania. While `pcm_zcr_sma` remains important, its contribution is slightly more pronounced for mania than depression. Features such as `pcm_fftMag_spectralcentroid_sma` and `slope0500_sma3` show a strong and specific impact on mania predictions, which is less evident in the depression plot. A key difference between the two states lies in how certain features are utilized. While `pcm_fftMag_mfcc_12_` is crucial for both states, its magnitude of impact is significantly greater for depression than for mania. On the other hand, `pcm_LOGenenergy_sma`, though important for depression, has a much more dominant influence in predicting mania.

When we look at next part about symptoms, in Figure 34 there are presented the overall importance of features across multiple symptoms. The top contributors include `pcm_zcr_sma`, `pcm_fftMag_mfcc_12_`, and `pcm_LOGenenergy_sma`, which significantly impact the model's predictions. The feature `pcm_zcr_sma` has the highest mean SHAP value, reflecting its dominant role in distinguishing between symptoms such as elevated mood, anxiety, and destruc-

tive behavior. The balanced color distribution across symptoms suggests that most features contribute to multiple symptoms, though their impact varies in magnitude.

The SHAP summary plots (depicted in Figure 35) for anxiety and elevated mood disclose different patterns in how features indicate their importance between different symptoms. For anxiety, the most influential features are `pcm_fftMag_mfcc_12_` and `pcm_zcr_sma`. These features show strong, where high values often push predictions towards anxiety. Features such as `pcm_LOGenenergy_sma` and `slope0-500_sma3` also play significant roles but with more moderate impacts compared to the top features. In contrast, elevated mood relies more heavily on features such as `pcm_fftMag_spectralRollOff90_0_sma`, `pcm_fftMag_mfcc_12_` and `slope0500_sma3`.

It is important to consider multiple approaches and to investigate the impact of label uncertainty on the final outcomes. Although determining the optimal set of features remains challenging, it is acknowledged that this uncertainty significantly influences the features, thus justifying further research.

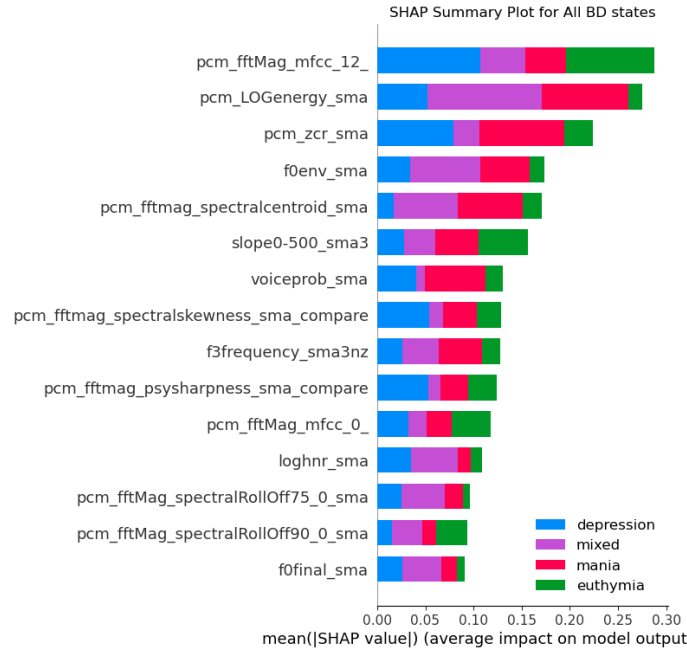


Figure 32: The SHAP summary plot indicates an influence of the most important features for all of the classes.

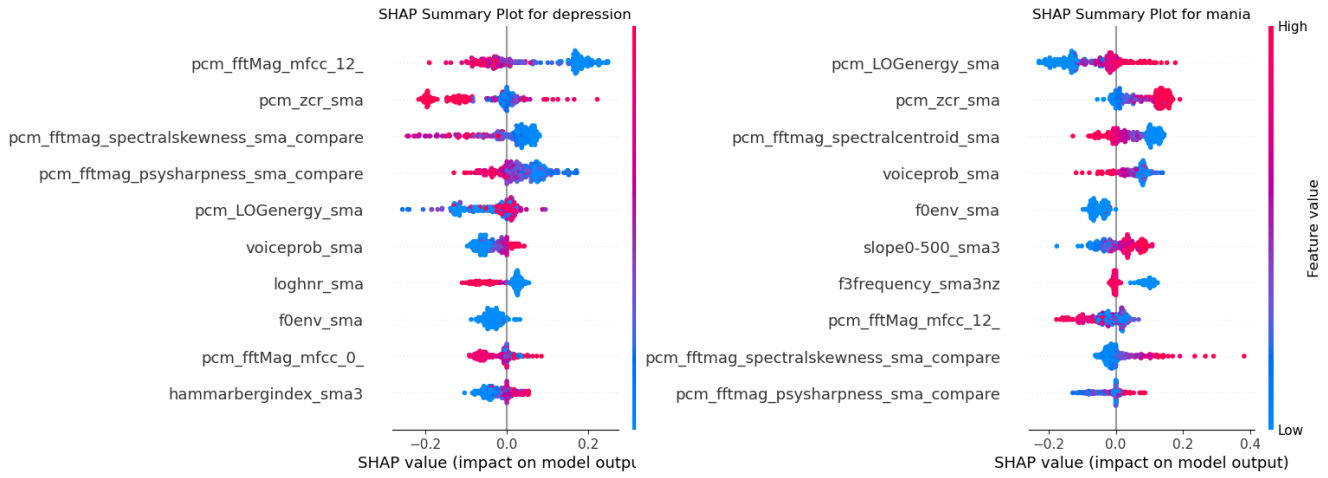


Figure 33: The SHAP summary plot indicates an influence of the most important features for depression (left) and mania (right)

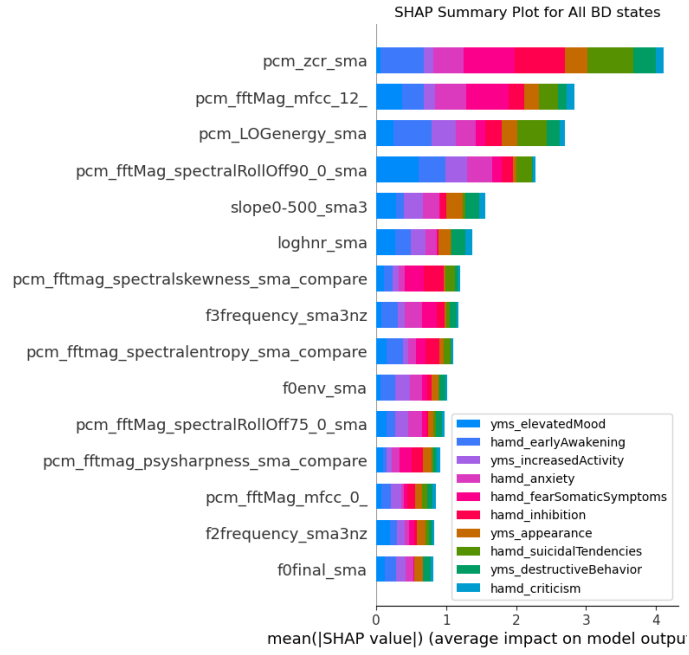


Figure 34: The SHAP summary plot indicates an influence of the most important features for all of the symptoms

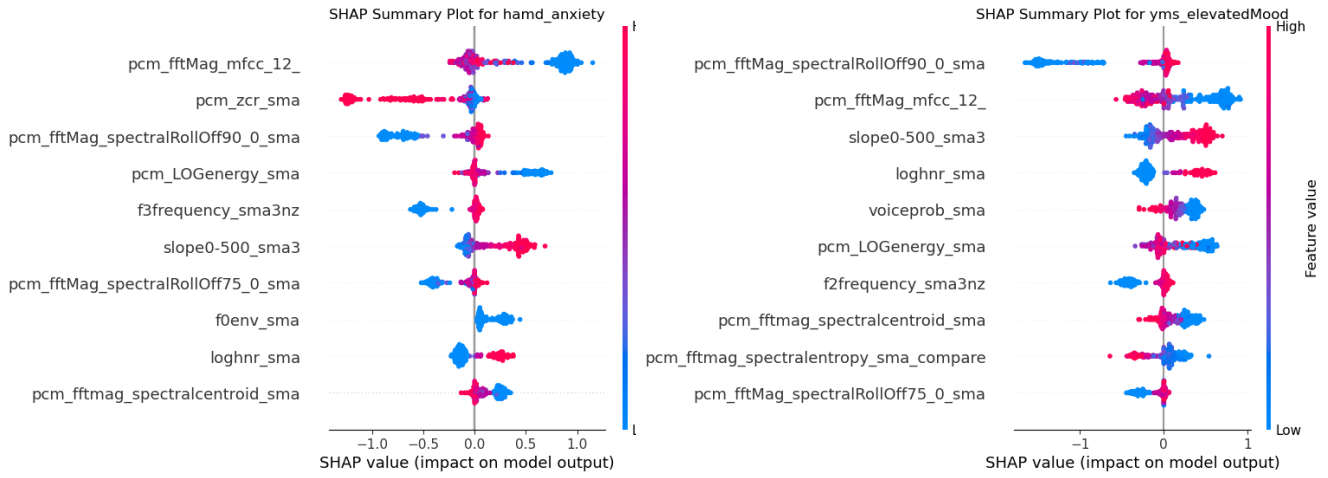


Figure 35: The SHAP summary plot indicates an influence of the most important features for anxiety (left) and elevated mood (right)

7.3.5 Semi-supervised Cluster Validation Agreement

Presented in Section 6, the Semi-supervised Cluster Validation Agreement (SCVA) method presents a validation approach for cluster consistency obtained during two different visits of patients with bipolar disorder.

The first step involves clustering all data collected in the study, where acoustic parameter values are aggregated by telephone calls and represented by the vector W_{rp} . Only those acoustic parameters identified as the best predictors in Section 4.2 were selected for clustering. In this dissertation, we consider 4 clustering methods using self-organizing maps. In Figure 36 we present results of clustering for a patient who has two visits assessed as Euthymia.

Next, Figure 37 is a second comparison of that patient with different assignments (mania and depression). The goal is to illustrate the clustering of patient data during two visits assessed concordant (E-E) and contrary (M-D) in particular patient. The X-axis represents the mobile recording date, and the Y-axis represents the cluster number assigned to each data point for each specific SOM configuration. The color gradient of the data points indicates the confidence level of assigning that data point to the corresponding cluster.

In Figure 36 first of all we observe that difference between number of mobile calls in those 2 time windows differ significantly. Therefore a detailed comparison may be impossible. In that case, only the best-preserved balance is viable across all clusters and presented in Topology 1x3. Next in Figure 37 we observe comparable number of recording between two classes. First topology 1x2 focuses all values into first cluster. Next topologies 1x4 and 2x2 grouped values into 3 cluster (instead of 4) what could suggest that values are similar. The best differentiation is visible in topology 1x3 where recordings are grouped into first cluster in Mania, and in Depression its distributed comparably into 3 clusters.

The heatmaps display the clustering validation agreement between two visits for the same patients, each labeled with a medical assessment, where: E (euthymia), M (mania), D (depression), and X (mixed state). The heatmaps compare different topologies (1x2, 1x3, 1x4, and 2x2), aiming to evaluate how consistently patients are classified in these conditions between two visits. Higher values on the diagonal represent better consistency for the same diagnosis, while lower values off the diagonal imply reduced misclassifications across different BD states.

In the first set of heatmaps (Figure 38), we can not observe notable trends in terms of diagonal and off-diagonal values. The diagonal values, which indicate agreement for the same diagnosis between visits, tend to be slightly higher for most topologies. The highest results for diagonal we observe for topology 1x2 and 2x2, whereas the lowest (demanding) values for off-diagonal we observe for topologies 1x3 and 1x4.

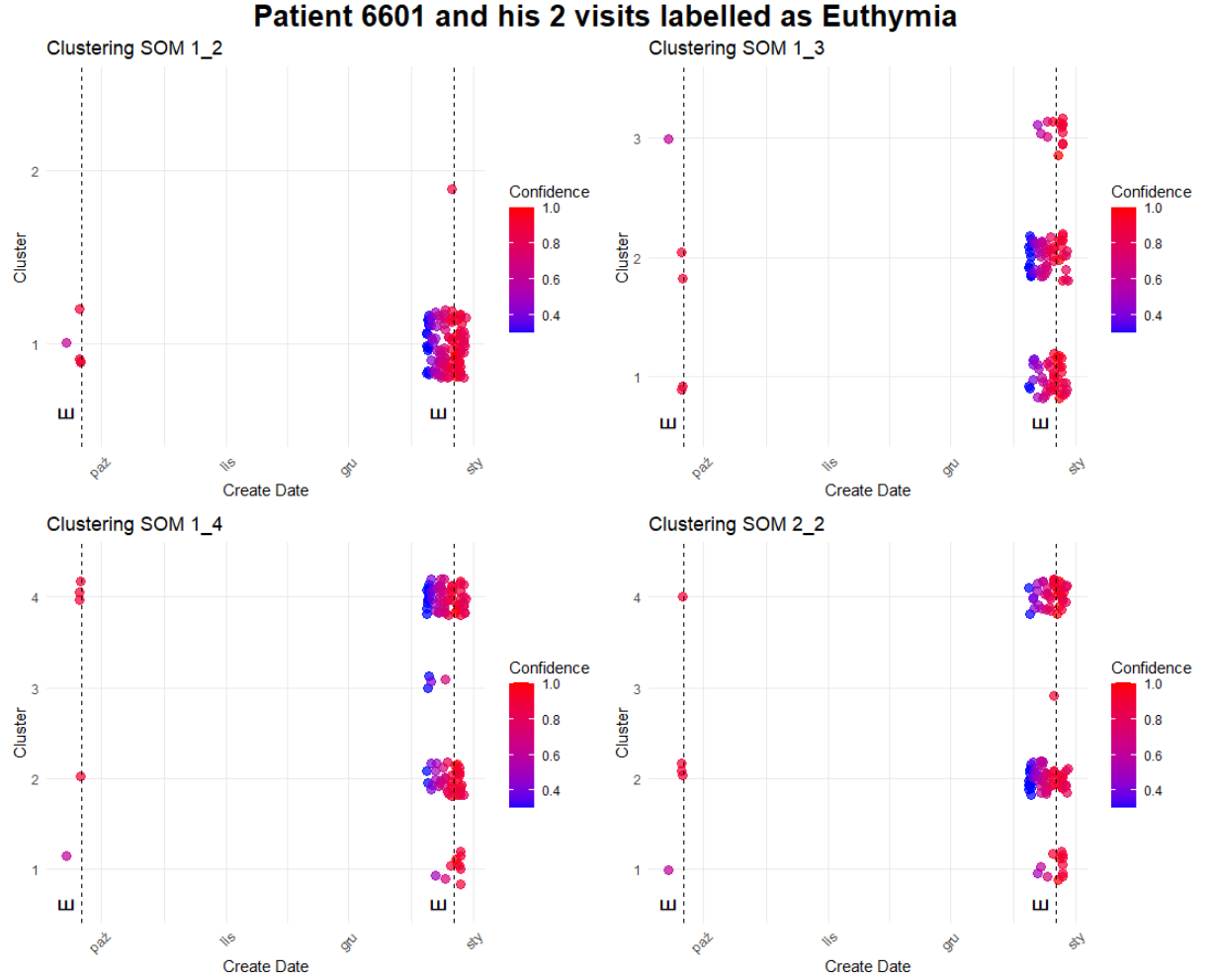


Figure 36: The considered four types of SOM topologies used during clustering and SCVA method where patient received 2 equal medical assessment (Euthymia) during 2 visits.

However, when we look at Table 16 in details, we can observe that number of available mobile recordings between two visits (columns Obs1 and Obs2 consecutively Visit Ratio) could vary markedly that significantly affects the interpretation and the results. The comparison between visits that one of them has several observation and next has few times more could be misleading. Therefore in Figure 39 we present another set of heatmaps selecting only values that received ratio of number of mobile recordings between two visit grater then 0.5. In that case, the differences between diagonal and off-diagonal become even clearer.

The most important observation is the lack of class X-X, which means that the compared values differed significantly from each other - which could have influenced

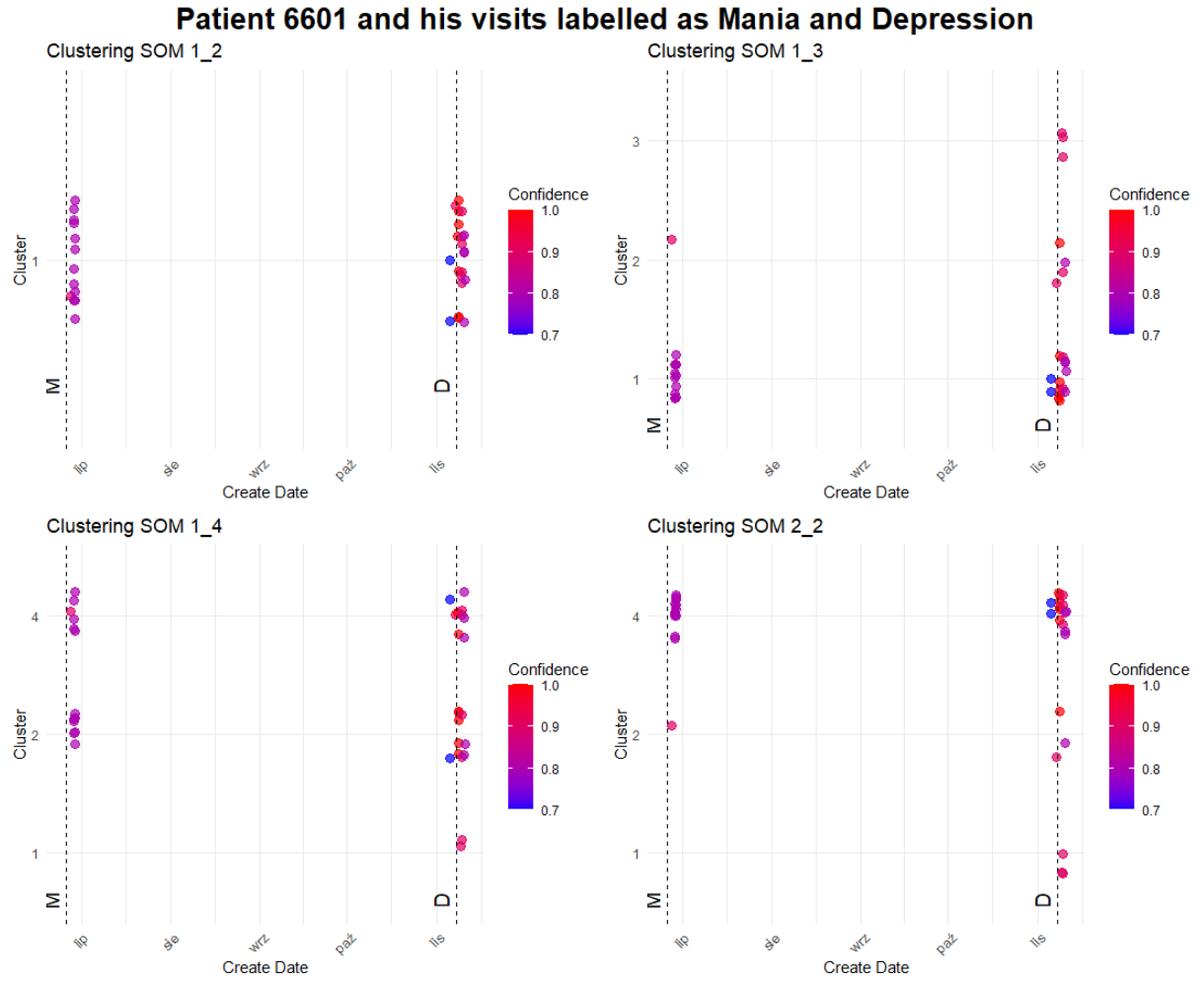


Figure 37: The considered four types of SOM topologies used during clustering and SCVA method where patient received 2 different medical assessment (Mania and Depression) during 2 visits.

the previous result.

Between all of topologies the best results were obtained for topology 1x4 - it means that difference between mean of diagonal across all states and mean of off-diagonal across all states is the highest and equal 1.0.

Focusing on that topology we observe that values on that diagonal are highest across all states. In particular Mania we observe 0.71 between 2 visits labeled as Mania where 0.64 between Mania and Euthymia and 0.57 for Mania - Depression, what is significant because it describes the two state we very the most in real life. Comparing SCVA values between euthymia and the others there is well-illustrating difference. In E-E we observe the highest values, E-X is the lowest what reflects

well the real condition of patients.

In the remaining topologies the worst results are in Topology 1x2 due to relatively high value across all of combination. In topology 1x3 we see the smallest value of all maps equal to 0.47 for the combination M - X, however we can also see the value 0.63 which is comparable to the values on the diagonal. In topology 2x2 all the values on the diagonal are above 0.7 but the rest values do not differ much and are between 0.58 and 0.67.

Details results on which the heatmaps were built are presented in Table 16. That summary table consist of list of considered patients and combination of visits using SCVA method.

The table presents data where each row corresponds to a specific combination of patient visits, where columns are as follow:

PID: patient ID, VComb: combination of patient visit, V1,V2: date of visit 1 and 2, L1 and L2: psychiatric assignment (label 0:Euthymia, 1:Depression, 2:Mania, 3:Mixed), Obs1, Obs2: number of collected observation around Visit 1 and 2, Visit Ration: ratio between 2 number of collected observation around visits, SCVA: result for SCVA for 4 considered topologies.

In addition to the data structure, several key observations and interpretations can be made. Starting with column called Visit Ratio that reflects the balance between these observation counts—values close to 1 indicate similar numbers of observations at both visits, while lower values reveal a larger discrepancy. First, the existence of multiple visit combinations for some patients allows for a comparative analysis of how different time intervals and varying numbers of observations influence the SCVA outcomes. For instance, patient 681 has 85 observations at the first visit and 76 at the second, yielding a Visit Ratio of approximately 0.89, that may contribute to pay higher attention to this data. In contrast, patient 6139 shows a significant imbalance with 62 observations at the first visit versus only 11 at the second, resulting in a Visit Ratio of about 0.18 suggest high disproportion of observation in that period.

Moreover the variability, combined with the differences in observation counts, offers insights into which conditions or topologies yield more reliable results. For example, in patient 6701, the combination of visits 1 and 2 results in a very high SCVA score of 0.92 (topology 1.4) and Visit Ration equal 0.97, which is consistent with the patient's BD state.

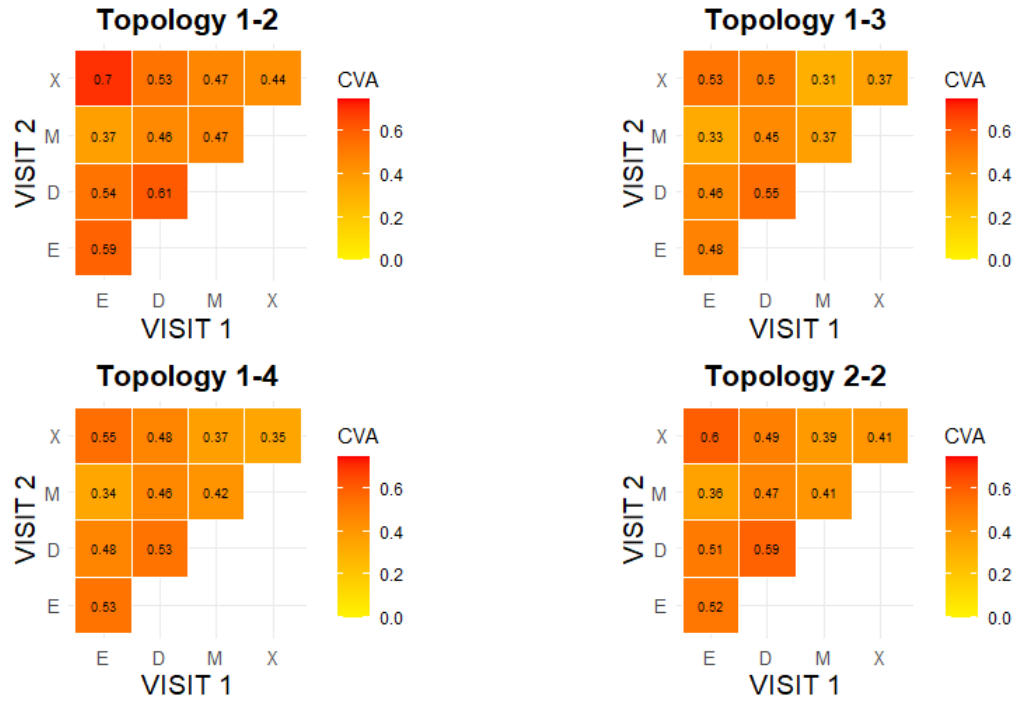


Figure 38: The heatmaps for considered pair of 2 visits for different types of topologies with result of SCVA value.

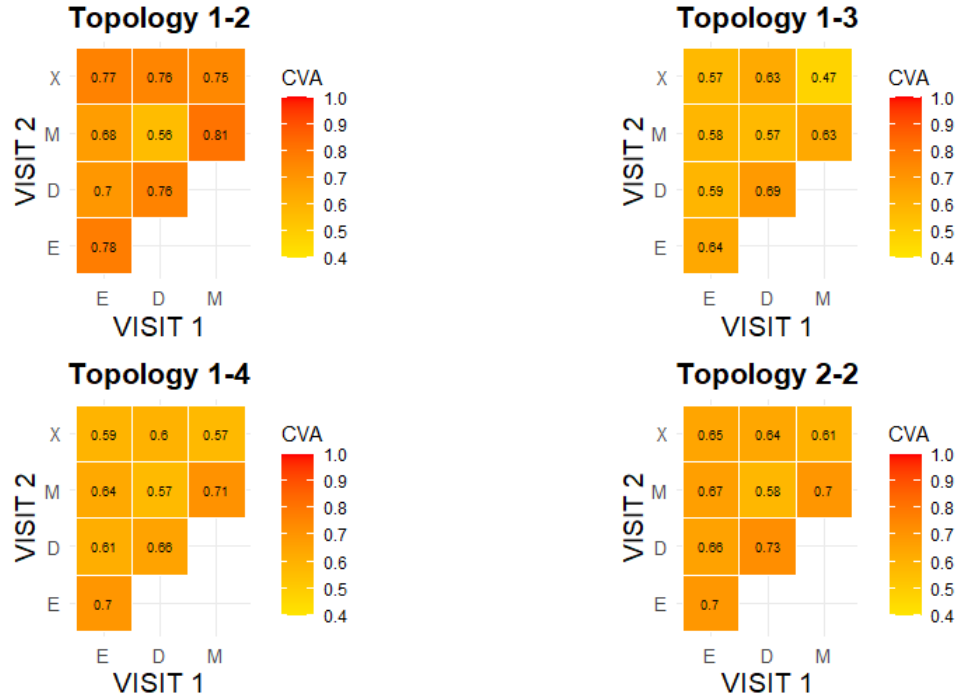


Figure 39: The heatmaps for considered pair of 2 visits for different types of topologies with result of SCVA value where parameter `visit_ratio` is above 0.5

Table 16: The summary table with list of considered patients and combination of visits using SCVA method.

PID: patient ID, VComb: combination of patient visit, V1,V2: date of visit 1 and 2, L1 and L2: psychiatric assignment (label 0:Euthymia, 1:Depression, 2:Mania, 3:Mixed), Obs1, Obs2: number of collected observation around Visit 1 and 2, Visit Ratio: ratio between 2 number of collected observation around visits, SCVA: result for SCVA for 4 considered topologies.

PID	VComb	V1	V2	L1	L2	Obs1	Obs2	Visit Ratio	SCVA 1.2	SCVA 1.3	SCVA 1.4	SCVA 2.2
681	1 - 2	2018-08-23	2018-11-10	1	1	85	76	0.89	0.91	0.79	0.73	0.82
1153	1 - 2	2018-08-02	2018-11-02	1	1	6	7	0.86	0.71	0.58	0.62	0.68
1153	1 - 3	2018-08-02	2019-01-24	1	0	6	6	1.00	0.68	0.61	0.67	0.78
1153	2 - 3	2018-11-02	2019-01-24	1	0	7	6	0.86	0.61	0.55	0.66	0.67
1235	1 - 2	2018-11-09	2018-11-29	1	1	6	9	0.67	0.73	0.76	0.72	0.77
1472	1 - 2	2018-03-28	2018-06-20	3	0	188	142	0.76	0.85	0.67	0.68	0.72
1472	1 - 3	2018-03-28	2018-08-07	3	3	188	73	0.39	0.56	0.44	0.35	0.46
1472	2 - 3	2018-06-20	2018-08-07	0	3	142	73	0.51	0.65	0.44	0.45	0.52
1981	1 - 2	2018-08-09	2018-11-09	2	0	21	109	0.19	0.36	0.25	0.22	0.30
2004	1 - 2	2018-04-18	2018-07-19	0	0	20	147	0.14	0.24	0.18	0.23	0.18
2004	1 - 3	2018-04-18	2018-08-02	0	1	20	70	0.29	0.43	0.23	0.29	0.31
2004	1 - 3	2018-07-19	2018-08-02	0	1	147	70	0.48	0.57	0.48	0.51	0.47
2004	2 - 3	2018-07-19	2018-08-02	0	1	147	70	0.48	0.57	0.48	0.51	0.47
2500	1 - 2	2018-08-02	2018-09-03	2	2	11	124	0.09	0.23	0.23	0.25	0.25
2582	1 - 2	2018-06-19	2018-07-17	0	3	57	75	0.76	0.83	0.56	0.61	0.65
2582	1 - 3	2018-06-19	2018-09-28	0	0	57	91	0.63	0.68	0.44	0.47	0.52
2582	1 - 4	2018-06-19	2018-10-09	0	1	57	69	0.83	0.71	0.58	0.54	0.60
2582	2 - 3	2018-07-17	2018-09-28	3	0	75	91	0.82	0.79	0.59	0.48	0.62
2582	2 - 4	2018-07-17	2018-10-09	3	1	75	69	0.92	0.77	0.66	0.50	0.61
2582	2 - 5	2018-05-22	2018-08-29	1	2	84	31	0.37	0.40	0.35	0.38	0.42
2582	3 - 4	2018-09-28	2018-10-09	0	1	91	69	0.76	0.65	0.59	0.61	0.61
3839	1 - 2	2018-08-07	2018-11-14	1	0	9	18	0.50	0.54	0.45	0.51	0.51
4157	1 - 2	2018-09-17	2018-12-17	0	1	19	3	0.16	0.26	0.20	0.25	0.25
4248	1 - 2	2018-03-27	2018-05-22	1	1	44	84	0.52	0.72	0.67	0.68	0.72
4248	1 - 3	2018-03-27	2018-06-20	1	1	44	115	0.38	0.50	0.41	0.42	0.47
4248	1 - 3	2018-06-26	2018-11-08	2	1	13	21	0.62	0.65	0.50	0.55	0.55
4248	1 - 4	2018-03-27	2018-07-30	1	1	44	82	0.54	0.61	0.50	0.48	0.56
4248	1 - 5	2018-03-27	2018-08-29	1	2	44	31	0.70	0.55	0.53	0.55	0.58
4248	2 - 3	2018-05-22	2018-06-20	1	1	84	115	0.73	0.69	0.60	0.58	0.64
4248	2 - 4	2018-05-22	2018-07-30	1	1	84	82	0.98	0.83	0.69	0.66	0.76
4248	2 - 5	2018-05-22	2018-08-29	1	2	84	31	0.37	0.40	0.35	0.38	0.42
4248	3 - 4	2018-06-20	2018-07-30	1	1	115	82	0.71	0.81	0.72	0.69	0.77
4248	3 - 5	2018-06-20	2018-08-29	1	2	115	31	0.27	0.33	0.30	0.37	0.34
4248	4 - 5	2018-07-30	2018-08-29	1	2	82	31	0.38	0.37	0.32	0.36	0.38
5656	1 - 2	2018-06-07	2018-06-22	0	0	6	10	0.60	0.61	0.43	0.55	0.55
5656	1 - 3	2018-06-07	2018-07-10	0	1	6	12	0.50	0.53	0.48	0.47	0.54
5656	1 - 4	2018-06-07	2018-07-30	0	1	6	4	0.67	0.58	0.48	0.57	0.56
5656	2 - 3	2018-06-22	2018-07-10	0	1	10	12	0.83	0.72	0.52	0.61	0.68
5656	2 - 4	2018-06-22	2018-07-30	0	1	10	4	0.40	0.43	0.35	0.45	0.46
5656	3 - 4	2018-07-10	2018-07-30	1	1	12	4	0.33	0.41	0.40	0.39	0.44
5854	1 - 2	2018-09-14	2018-10-04	2	1	20	45	0.44	0.50	0.44	0.45	0.47
5854	1 - 3	2018-09-14	2018-10-19	2	1	20	25	0.80	0.57	0.65	0.69	0.63
5854	1 - 4	2018-09-14	2018-12-14	2	1	20	35	0.57	0.55	0.62	0.56	0.61
5854	2 - 3	2018-10-04	2018-10-19	1	1	45	25	0.56	0.61	0.58	0.58	0.57
5854	2 - 4	2018-10-04	2018-12-14	1	1	45	35	0.78	0.83	0.70	0.70	0.75
5854	3 - 4	2018-10-19	2018-12-14	1	1	25	35	0.71	0.68	0.66	0.71	0.66
6139	1 - 2	2018-05-23	2018-07-24	2	0	62	11	0.18	0.23	0.29	0.27	0.27
6139	1 - 3	2018-05-23	2018-10-03	2	1	62	29	0.47	0.53	0.64	0.59	0.57
6139	1 - 4	2018-05-23	2018-11-13	2	1	62	88	0.70	0.48	0.57	0.49	0.53
6139	2 - 3	2018-07-24	2018-10-03	0	1	11	29	0.38	0.44	0.45	0.39	0.44
6139	2 - 4	2018-07-24	2018-11-13	0	1	11	88	0.13	0.26	0.25	0.20	0.22
6139	3 - 4	2018-10-03	2018-11-13	1	1	29	88	0.33	0.46	0.42	0.38	0.45

PID	VComb	V1	V2	L1	L2	Obs1	Obs2	Visit Ratio	SCVA 1.2	SCVA 1.3	SCVA 1.4	SCVA 2.2
6601	1 - 2	2018-06-26	2018-09-26	2	0	13	5	0.38	0.51	0.48	0.48	0.51
6601	1 - 3	2018-06-26	2018-11-08	2	1	13	21	0.62	0.65	0.50	0.55	0.55
6601	1 - 4	2018-06-26	2018-12-27	2	0	13	111	0.12	0.20	0.09	0.12	0.14
6601	2 - 3	2018-09-26	2018-11-08	0	1	5	21	0.24	0.35	0.27	0.30	0.32
6601	2 - 4	2018-09-26	2018-12-27	0	0	5	111	0.05	0.14	0.07	0.07	0.09
6601	3 - 4	2018-11-08	2018-12-27	1	0	21	111	0.19	0.30	0.26	0.26	0.29
6701	1 - 2	2018-08-02	2018-08-07	0	0	90	93	0.97	0.98	0.84	0.92	0.88
6701	1 - 3	2018-08-02	2018-08-24	0	0	90	53	0.59	0.77	0.65	0.71	0.68
6701	1 - 4	2018-08-02	2018-11-08	0	0	90	246	0.37	0.50	0.37	0.46	0.43
6701	2 - 3	2018-08-07	2018-08-24	0	0	93	53	0.57	0.75	0.75	0.75	0.75
6701	2 - 4	2018-08-07	2018-11-08	0	0	93	246	0.38	0.51	0.47	0.51	0.51
6701	3 - 4	2018-08-24	2018-11-08	0	0	53	246	0.22	0.38	0.35	0.38	0.38
6754	1 - 2	2018-07-18	2018-09-29	0	2	1	7	0.14	0.25	0.28	0.31	0.28
7501	1 - 2	2018-07-25	2018-08-21	2	2	12	62	0.19	0.36	0.29	0.34	0.33
7501	1 - 3	2018-07-25	2018-09-12	2	2	12	61	0.20	0.37	0.27	0.34	0.31
7501	1 - 4	2018-07-25	2018-10-25	2	2	12	100	0.12	0.28	0.18	0.21	0.21
7501	2 - 3	2018-08-16	2018-09-12	2	2	62	61	0.98	0.94	0.80	0.94	0.88
7501	2 - 4	2018-08-16	2018-10-25	2	2	62	100	0.62	0.76	0.52	0.59	0.60
7501	3 - 4	2018-09-12	2018-10-25	2	2	61	100	0.61	0.74	0.59	0.61	0.62
8292	1 - 2	2018-06-11	2018-07-20	0	0	76	90	0.84	0.90	0.71	0.80	0.80
8779	1 - 2	2018-08-27	2018-11-27	1	0	6	17	0.35	0.33	0.44	0.44	0.45
8866	1 - 2	2018-07-03	2018-08-21	1	3	7	17	0.41	0.46	0.52	0.56	0.48
8866	1 - 3	2018-07-03	2018-09-03	1	1	7	5	0.71	0.60	0.63	0.60	0.69
8866	2 - 3	2018-08-21	2018-09-03	3	1	17	5	0.29	0.37	0.47	0.35	0.43
9341	1 - 2	2018-05-28	2018-06-04	3	2	1	7	0.14	0.26	0.22	0.26	0.26
9341	1 - 3	2018-05-28	2018-06-22	3	3	1	5	0.20	0.32	0.29	0.36	0.36
9341	1 - 4	2018-05-28	2018-07-20	3	0	1	6	0.17	0.33	0.27	0.30	0.30
9341	2 - 3	2018-06-04	2018-06-22	2	3	7	5	0.71	0.63	0.49	0.58	0.61
9341	2 - 4	2018-06-04	2018-07-20	2	0	7	6	0.86	0.68	0.58	0.64	0.67
9341	3 - 4	2018-06-22	2018-07-20	3	0	5	6	0.83	0.60	0.48	0.64	0.64
9813	1 - 2	2018-10-10	2018-10-30	0	1	34	199	0.17	0.30	0.19	0.21	0.23
9813	1 - 3	2018-10-10	2018-12-06	0	3	34	31	0.91	0.87	0.69	0.71	0.75
9813	1 - 4	2018-10-10	2019-01-24	0	1	34	38	0.89	0.69	0.44	0.54	0.57
9813	2 - 3	2018-10-30	2018-12-06	1	3	199	31	0.16	0.32	0.22	0.27	0.25
9813	2 - 4	2018-10-30	2019-01-24	1	1	199	38	0.19	0.42	0.25	0.34	0.30
9813	3 - 4	2018-12-06	2019-01-24	3	1	31	38	0.82	0.74	0.60	0.70	0.68
9829	1 - 2	2018-07-03	2018-07-24	1	1	4	18	0.22	0.35	0.35	0.32	0.38
9829	1 - 3	2018-07-03	2018-09-05	1	1	4	32	0.13	0.27	0.27	0.26	0.29
9829	1 - 4	2018-07-03	2018-10-01	1	1	4	18	0.22	0.35	0.39	0.36	0.41
9829	1 - 5	2018-07-03	2018-10-22	1	1	4	22	0.18	0.31	0.32	0.32	0.35
9829	1 - 6	2018-07-03	2018-11-13	1	1	4	52	0.08	0.20	0.20	0.20	0.22
9829	2 - 3	2018-07-24	2018-09-05	1	1	18	32	0.56	0.67	0.59	0.56	0.63
9829	2 - 4	2018-07-24	2018-10-01	1	1	18	18	1.00	0.89	0.85	0.77	0.88
9829	2 - 5	2018-07-24	2018-10-22	1	1	18	22	0.82	0.82	0.83	0.68	0.85
9829	2 - 6	2018-07-24	2018-11-13	1	1	18	52	0.35	0.54	0.46	0.45	0.50
9829	3 - 4	2018-09-05	2018-10-01	1	1	32	18	0.56	0.70	0.65	0.61	0.68
9829	3 - 5	2018-09-05	2018-10-22	1	1	32	22	0.69	0.78	0.66	0.64	0.70
9829	3 - 6	2018-09-05	2018-11-13	1	1	32	52	0.62	0.77	0.65	0.66	0.69
9829	4 - 5	2018-10-01	2018-10-22	1	1	18	22	0.82	0.82	0.78	0.71	0.81

7.4 Discussion

The analysis undertaken in this study highlights several significant challenges inherent in processing complex datasets within the context of bipolar disorder. The data were notably difficult to interpret due to substantial gaps and a limited number of labels. These deficiencies constrained the selection of analytical methods and underscored the importance of meticulous data preparation and feature selection.

Our approach to feature selection, though not widely adopted in existing literature, revealed that not all parameters contribute meaningfully to the classification task. By identifying and focusing on the most relevant features, we streamlined the data preparation process, which is crucial for handling large and intricate datasets. We observed that the selection of important features differs in particular BD states and should not be ignored with modeling in that disease. The application of the Cost-Constrained Feature Selection test further demonstrated the superiority of acoustic features over basic, easily collectible parameters such as the daily number of calls. This finding emphasizes the value of in-depth feature engineering, particularly when subtle nuances in data can significantly impact outcomes.

Additionally, the study uncovered a pressing need for more comprehensive labeling strategies. The current scarcity and ambiguity of labels—notably the lack of granularity regarding degrees of certainty—limit the effectiveness of model training and validation. Incorporating labels with varying degrees of confidence could enhance model robustness and provide more nuanced insights into patient states, which is particularly important in medical contexts where ambiguity is common.

The choice of neural network architectures and their configurations was found to influence classification accuracy significantly. Our experiments demonstrated noticeable differences between data corresponding to different states of bipolar disorder, confirming the potential of acoustic features in distinguishing these states. Selecting an appropriate architecture is thus not merely a technical consideration but a crucial factor that can affect the applicability of the model in real-world settings.

It is important to recognize that classification performance alone does not suffice in the context of medical AI applications. Recent research trends emphasize the importance of the model itself—how well it aligns with the problem domain and its interpretability. A model that fits the specific nuances of BD and can be understood by medical professionals holds greater significance than one that simply achieves high accuracy metrics.

Clustering unlabeled data presented its own set of challenges. The absence of a clear, optimal clustering topology for bipolar disorder data indicates that multiple approaches may be viable, each with its own trade-offs. The clusters generated did not directly correspond to specific patient states, highlighting the necessity of validating cluster assignments against available labeled data while accounting for

label uncertainty. This step is critical to ensure that the clusters are meaningful and can be interpreted in a clinically relevant manner.

Focusing on the detailed description and analysis of acoustic features has a profound impact on the overall solution, especially in the application of AI in medicine. Acoustic features can capture subtle changes in patient behavior and state, which are often not detectable through basic parameters. This focus enhances the model's ability to provide clinically relevant insights, thereby improving the potential for timely and accurate interventions.

8 Conclusions

The inspiration behind this research on datasets collected from bipolar disorder patients was primarily the opportunity to model and develop various innovative methods. Bipolar disorder presents numerous aspects that require careful consideration during research preparation. A significant example is the bimodal nature of the disease—it manifests with features of depression as well as entirely different symptoms of mania.

An important issue, often overlooked in other studies, is the approach to handling the label uncertainty arising during their assignment by psychiatrists. Since patients are not hospitalized, the label assigned by the doctor is fully valid only for the day of the assessment. Therefore, extending the label should incorporate its associated uncertainty.

The main achievements of this doctoral dissertation can be summarized as follows:

1. The proposed **Uncertainty Aware Advanced Feature Selection** method by incorporating label uncertainty during the feature selection process enhances the robustness of the model. This approach acknowledges that labels assigned by clinicians may have inherent uncertainties due to the fluctuating nature of the disorder.
2. The proposed **Explainable Compositional Multi-Layer Perceptron with Confidence Factor** enables creating an interpretable model that can classify multiple outputs while considering label uncertainty.
3. The proposed **Semi-supervised Cluster Validation Agreement** accounting for label uncertainty improves the reliability of clustering results. This is crucial when dealing with unlabeled or partially labeled data common in medical contexts.

Extensive experiments confirmed that the three theses stated in the Introduction have been successfully achieved:

- (a) *The proposed Uncertainty Aware Advanced Feature Selection method effectively addresses the uncertainty arising from psychiatric label assignment over time, ensuring the selection of relevant and adequate acoustic features for the improved classification of bipolar disorder states.*

This work highlights the most important features for monitoring bipolar disorder patients and differentiates these features based on the disease states (depression and mania). Experiments confirmed that classification based on selected parameters achieves better results when label

uncertainty is simultaneously modeled. Specifically, using uncertainties that differ from a constant value (e.g., Step uncertainty or Gaussian distribution) generally yields better classification outcomes than assuming equal importance of all labels. In particular, the accuracy of BD classification formulated as binary problem (depression vs euthymia) accounted to 0.77 for UAAFS with Gaussian function representing label confidence and 0.73 for the constant function. Thus, we conclude that the proposed solution effectively addresses the main feature selection problem posed in this dissertation.

- (b) *The proposed Explainable Compositional Multi-Layer Perceptron with Confidence Factor (XCMLPCF) method enables to reflect the uncertainty related to affective symptoms adequately. The proposed two-layer and multi-output approach enables both prediction and explanation of features that impact the resulting state and symptoms, which is particularly important in the medical field.*

Furthermore for **XCMLPCF** several neural network architectures were tested, yielding varying classification results. Most importantly, a framework was developed that enables a comprehensive approach to the problem, highlighting the application of neural networks along with their interpretation and the incorporation of appropriate uncertainty associated with medical labels. The prepared graphs clearly indicate which features most affect the final result, which is very much needed in the medical industry. Additionally, implementing an architecture that consists of multiple outputs helps better understand the patient's current state.

- (c) *The proposed Semi-supervised Cluster Validation Agreement (SCVA) enables to effectively determine the degree of similarity of unlabeled data to labeled data considering the temporal nature of sensor and psychiatric data collection. Thus, it is a promising approach to support the model validation process.*

The **SCVA** method determines the degree of similarity between labeled and unlabeled data, considering the distance from a given label. Four different topologies for initial data clustering were considered. The approach with four classes (in accordance with the nature of the data) achieved the best similarity results – an average of 0.69 for groups with similar labels and 0.59 for groups that differ from each other. Experimental validation of the obtained clusters clearly indicated the need for diverse approaches to the problem. The choice of problem topology is non-obvious, and similarities between particular states especially between depression and euthymia may be influenced by the size of the

data sets. Incorporating label uncertainty models the problem more effectively than previous solutions what is a the solution to the secondary problem posed in this dissertation.

The developed general scheme for classifying states in bipolar disorder captures multiple dimensions of the disease. Beyond basic state classification, introducing the classification of specific symptoms provides valuable insights for psychiatrists. By incorporating Explainable AI, clinicians can observe which features are significant during selected phone calls over time, along with their modeled uncertainty. Preliminary experiments indicate that the choice of neural network architecture is crucial—the traditional Artificial Neural Networks achieve worse results compared to Convolutional Neural Networks, which are particularly effective for acoustic data. The proposed solution is a continuation of solving the main problem of this dissertation.

In addition to my dissertation work, I actively engaged in various projects and tasks that enriched my overall research experience. For example, I received travel funding and had the opportunity to present a scientific poster about our work [48] during Women in Machine Learning Workshop 2022 in the USA during the NeurIPS Conference, which not only showcased my work on innovative methodologies but also allowed me to network with international experts. I further expanded my academic horizons by presenting my doctoral research at the University in Vienna in December 2024, where I became a laureate during competition called 'Young Researchers in Vienna' organized by Scientific Center of the Polish Academy of Sciences in Vienna, facilitating an exchange of ideas with a broader academic community.

Moreover, I contributed to a clinical bipolar project called "Moodmon" where I was responsible for creating the analytical package that supported our research efforts. My involvement in this project provided practical insights into clinical applications and data analysis. Additionally, professional work at Britenet Company allowed me to develop further research on bipolar disorder, where our team explored advanced techniques and collaborative research avenues.

It is important to note that while several other projects were conducted, they were not included in the main body of this work due to their differing objectives or space limitations. Nonetheless, these experiences have significantly contributed to my professional development and broadened my perspective on the subject.

To conclude, this dissertation contributes novel methods for handling label uncertainty in the classification and clustering of various sensor and psychiatric data. By addressing the complexities inherent in the disease such as

its bimodal nature and the challenges in obtaining precise labels - the study advances the field toward more accurate and interpretable models. The integration of XAI techniques ensures that the models are not only effective but also transparent, which is crucial for clinical applications.

8.1 Future work

My future work will develop in several directions. First of all, I plan to focus on further improving the classification results of states and symptoms in bipolar disorder. With the rapid advancement of neural networks, there is a vast area for research that could lead to significantly better results. Additionally, treating data unsupervised is extremely important due to the difficulty in continuously obtaining accurate labels. Developing unsupervised approaches will enable easier and faster detection of mood changes in patients.

Secondly, my forthcoming work will be engaged with the 'Moodmon' clinical project, where I plan to continue model development. The implementation of the neural networks I tested and their adaptation to the problem addressed in the project may improve the results so far. However, an attempt to explain the influence of variables on the result will certainly prove helpful to the psychiatrist. Moreover, another challenge in that project could be model development for each patient independently. Methods like transfer learning could be helpful with this approach.

Next, further research will be devoted to explainability and linguistic summarization as a consecutive step to explain uncertainty-aware black box models. Applications in various domains have improved the understanding of large datasets using natural language statements [72]. Further work will leverage XAI system outputs to provide non-technicians, such as medical experts, with linguistic summaries that explain the relationship between feature importance and model outputs. As presented in [48], we described the relationship between input features and class predictions, indicating both positive and negative contributions based on Shapley values. For example, records that positively predict the depression class often have low-level voice quality features. Moreover, some features presented with the SHAP plots are much more important for the domain expert. Ideally, such preferences of experts towards features should be wisely included early in the modeling process. Future work will consider linguistic summarization as an extension of the proposed XCMLP-CF algorithm.

Furthermore, the degree of labeling depending on the patient's disease state will be addressed. Literature indicates that dynamics in mania and depression may differ. Thus, the confidence of the label extrapolation may also depend on the state.

Additionally, other mental disorders, such as anxiety disorders and schizophrenia, may need to implement label uncertainty due to the different duration of disease symptoms. This observation underscores the vast potential for applying these methods in diverse areas of medicine where uncertain labeling presents a significant challenge. Finally, the proposed methods can be easily extended to handle other application contexts.

References

- [1] Niloufar Abaei and Hussein Al Osman. “A Hybrid Model for Bipolar Disorder Classification from Visual Information”. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 4107–4111.
- [2] Niloufar AbaeiKoupaei and Hussein Al Osman. “A Multi-Modal Stacked Ensemble Model for Bipolar Disorder Classification”. In: *IEEE Transactions on Affective Computing* 14.1 (2023), pp. 236–244.
- [3] Amina Adadi and Mohammed Berrada. “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)”. In: *IEEE Access* 6 (2018), pp. 52138–52160.
- [4] Jorge Alvarez-Lozano et al. “Tell me your apps and I will tell you your mood: correlation of apps usage with bipolar disorder state”. In: *Proceedings of the 7th International Conference on Pervasive Technologies Related to Assistive Environments. PETRA '14*. Rhodes, Greece: Association for Computing Machinery, 2014.
- [5] Anna Z Antosik-Wojcinska et al. “Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling”. In: *Int J Med Inform* 138:104131 (2020).
- [6] Armen C. Arevian et al. “Clinical state tracking in serious mental illness through computational analysis of speech”. In: *PLOS ONE* 15.1 (Jan. 2020), pp. 1–17.
- [7] Jakob E. Bardram et al. “The MONARCA self-assessment system: a persuasive personal monitoring system for bipolar patients”. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. IHI '12*. Miami, Florida, USA: Association for Computing Machinery, 2012, 21–30.
- [8] Michael Bauer and Andrea Pfennig. “Epidemiology of bipolar disorders”. In: *Epilepsia* 46.SUPPL. 4 (2005), 8 – 13.
- [9] Till Beiwinkel et al. “Using Smartphones to Monitor Bipolar Disorder Symptoms: A Pilot Study”. In: *JMIR Mental Health* 3.1 (2016), e2.
- [10] M.B. Bonsall et al. “Nonlinear time-series approaches in characterizing mood stability and mood in stability in bipolar disorder”. In: *Proceedings of the Royal Society of London, Series B: Biological Sciences* 279 (2012), pp. 916–924.

- [11] Jedediah M Bopp et al. “The longitudinal course of bipolar disorder as revealed through weekly text messaging: a feasibility study”. In: *Bipolar Disorders* 12.3 (2010), pp. 327–334.
- [12] Bokai Cao et al. “DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection”. In: (Aug. 2017), pp. 747–755.
- [13] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, 2010.
- [14] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. ACM, Aug. 2016.
- [15] Laura Vergel de Dios Vicki Wang Colin A. Depp Daniel H. Kim and Jennifer Ceglowski. “A Pilot Study of Mood Ratings Captured by Mobile Phone Versus Paper-and-Pencil Mood Charts in Bipolar Disorder”. In: *Journal of Dual Diagnosis* 8.4 (2012), pp. 326–332.
- [16] Nicholas Cummins et al. “A review of depression and suicide risk assessment using speech analysis”. In: *Speech Communication* 71 (2015), pp. 10–49.
- [17] Nicholas Cummins et al. “Multilingual markers of depression in remotely collected speech samples: A preliminary analysis”. In: *Journal of Affective Disorders* 341 (2023), pp. 128–136.
- [18] Colin A Depp et al. “Augmenting psychoeducation with a mobile intervention for bipolar disorder: A randomized controlled trial”. In: *Journal of Affective Disorders* 174 (2015), pp. 23–30.
- [19] Low DM, Bentley KH, and Ghosh SS. Kessing. “Automated assessment of psychiatric disorders using speech: A systematic review”. In: *Laryngoscope Investig Otolaryngol* 31;5(1) (2020), pp. 96–116.
- [20] Yizhuo Dong and Xinyu Yang. “A hierarchical depression detection model based on vocal and emotional cues”. In: *Neurocomputing* 441 (2021), pp. 279–290.
- [21] Turker Tekin Erguzel, Gokben Hizli Sayar, and Nevzat Tarhan. “Artificial intelligence approach to classify unipolar and bipolar depressive disorders”. In: *Neural Computing and Applications* 27.6 (Aug. 2016), pp. 1607–1616.

- [22] Caroline Wanderley Espinola et al. “Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study”. In: *Research on Biomedical Engineering* 37 (1 Mar. 2021), pp. 53–64.
- [23] Florian Eyben et al. “Recent developments in opensmile, the munich open-source multimedia feature extractor”. In: *Proc. of the 21st ACM Int. Conf. on Multimedia*. 2013, pp. 835–838.
- [24] Maria Faurholt-Jepsen et al. “Behavioral activities collected through smartphones and the association with illness activity in bipolar disorder”. In: *Int J Methods Psychiatr Res.* 25(4) (2016), pp. 309–323.
- [25] Maria Faurholt-Jepsen et al. “Objective smartphone data as a potential diagnostic marker of bipolar disorder”. In: *Australian & New Zealand Journal of Psychiatry* 53.2 (2019), pp. 119–128.
- [26] Maria Faurholt-Jepsen et al. “Using digital phenotyping to classify bipolar disorder and unipolar disorder – exploratory findings using machine learning models”. In: *European Neuropsychopharmacology* 81 (2024), pp. 12–19.
- [27] Maria Faurholt-Jepsen et al. “Voice analyses using smartphone-based data in patients with bipolar disorder, unaffected relatives and healthy control individuals, and during different affective states”. In: *International Journal of Bipolar Disorders* 9.1 (2021), pp. 1–13.
- [28] Maria Faurholt-Jepsen et al. “Voice analysis as an objective state marker in bipolar disorder”. In: *Translational psychiatry* 6.7 (2016), e856–e856.
- [29] Jerome Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *The Annals of Statistics* 29 (Nov. 2000).
- [30] John Gideon, Emily Mower Provost, and Melvin Mcinnis. “Mood state prediction from speech of varying acoustic quality for individuals with bipolar disorder”. In: vol. 2016. Mar. 2016, pp. 2359–2363.
- [31] Germán González-Almagro et al. “Semi-Supervised Constrained Clustering: An In-Depth Overview, Ranked Taxonomy and Future Research Directions”. In: *arXiv preprint arXiv:2303.00522* (2023).
- [32] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. “Speech recognition with deep recurrent neural networks”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 6645–6649.

- [33] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. “Speech recognition with deep recurrent neural networks”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 6645–6649.
- [34] Agnes Grünerbl, Amir Muaremi, and Venet Osmani. “Smartphone-Based Recognition of States and State Changes in Bipolar Disorder Patients”. In: *IEEE Journal of Biomedical and Health Informatics* 19(1) (2015).
- [35] Andrea Guidi et al. “Features of vocal frequency contour and speech rhythm in bipolar disorder”. In: *Biomedical Signal Processing and Control* 37 (2017), pp. 23–31.
- [36] Isabelle Guyon and André Elisseeff. “An introduction to variable and feature selection”. In: *J. Mach. Learn. Res.* 3.null (2003), 1157–1182.
- [37] Isabelle Guyon et al. “Gene selection for cancer classification using support vector machines”. In: *Machine learning* 46.1-3 (2002), pp. 389–422.
- [38] Max HAMILTON. “Development of a Rating Scale for Primary Depressive Illness”. In: *British Journal of Social and Clinical Psychology* 6.4 (Dec. 1967), 278–296.
- [39] Shayan Hassantabar et al. “MHDeep: Mental Health Disorder Detection System Based on Wearable Sensors and Artificial Neural Networks”. In: *ACM Trans. Embed. Comput. Syst.* 21.6 (2022).
- [40] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. In: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY: Springer New York, 2009, pp. 101–137.
- [41] Diego Hidalgo-Mazzei et al. “Is a SIMPLe smartphone application capable of improving biological rhythms in bipolar disorder?” In: *Journal of Affective Disorders* 223 (2017), pp. 10–16.
- [42] Olgierd Hryniewicz, Katarzyna Kaczmarek-Majer, and Olga Kamińska. “Possibilistic method for the identification of a change-point for complex non-stationary time series”. In: *FUZZ-IEEE*. Vol. 832. series. publisher, 2021.
- [43] K. Izdebski. *Emotions in the Human Voice, Volume 3: Culture and Perception*. Emotions in the Human Voice t. 3. Plural Publishing, Incorporated, 2008.
- [44] A.K. Jain, Jianchang Mao, and K.M. Mohiuddin. “Artificial neural networks: a tutorial”. In: *Computer* 29.3 (1996), pp. 31–44.

- [45] J. Kacprzyk, R. R. Yager, and S. Zadrozny. “A Fuzzy Logic Based Approach to Linguistic Summaries of Databases”. In: *Journal of Applied Mathematics and Computer Science* 10 (2000), pp. 813–834.
- [46] Katarzyna Kacmarek and Monika Dominiak. “Acoustic Features from Speech as Markers of Depressive and Manic Symptoms in Bipolar Disorder: A Prospective Study”. In: *Acta Psychiatrica Scandinavica* (2024).
- [47] Katarzyna Kacmarek-Majer et al. “Acoustic features from speech as markers of depressive and manic symptoms in bipolar disorder: A prospective study”. In: *Acta Psychiatrica Scandinavica* n/a.n/a ().
- [48] Katarzyna Kacmarek-Majer et al. “PLENARY: Explaining black-box models in natural language through fuzzy linguistic summaries”. In: *Information Sciences* 614 (2022), pp. 374–399.
- [49] Olga Kaminska, Katarzyna Kacmarek-Majer, and Olgierd Hryniewicz. “Impact of clustering unlabeled data on classification: case study in bipolar disorder”. In: *Proceedings of the 17th Conference on Computer Science and Intelligence Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, D. Ślęzak (eds). ACSIS*. Vol. 30. 2022, 931–934.
- [50] Olga Kaminska, Tomasz Klonecki, and Katarzyna Kacmarek-Majer. “Feature Selection in Bipolar Disorder Episode Classification Using Cost-Constrained Methods”. In: *Explainable Artificial Intelligence and Process Mining Applications for Healthcare*. Ed. by Jose M. Juarez et al. Cham: Springer Nature Switzerland, 2024, pp. 36–40.
- [51] Olga Kamińska et al. “Self-organizing Maps Using Acoustic Features for Prediction of State Change in Bipolar Disorder”. In: *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems* (2020).
- [52] Dorota Kamińska et al. “The Role of Selected Speech Signal Characteristics in Discriminating Unipolar and Bipolar Disorders”. In: *Sensors* 24.14 (2024).
- [53] Olga Kamińska, Katarzyna Kacmarek-Majer, and Olgierd Hryniewicz. “Acoustic Feature Selection with Fuzzy Clustering, Self Organizing Maps and Psychiatric Assessments”. In: *Proceedings of Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU 2020, Lisbon* (2020).

- [54] Zahi N Karam et al. “Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech”. In: *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2014, pp. 4858–4862.
- [55] Justin Ker et al. “Deep Learning Applications in Medical Image Analysis”. In: *IEEE Access* 6 (2018), pp. 9375–9389.
- [56] Kamil Kmita et al. “Confidence path regularization for handling label uncertainty in semi-supervised learning: use case in bipolar disorder monitoring”. In: *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 2022, pp. 1–8.
- [57] Teuvo Kohonen. “Self-Organizing Maps”. In: (1995).
- [58] Tania Lecomte et al. “Mobile applications for mental health issues: A meta-review (Preprint)”. In: *JMIR mHealth and uHealth* 8 (Dec. 2019).
- [59] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.
- [60] Yu Lei et al. “A convolutional neural network-based diagnostic method using resting-state electroencephalograph signals for major depressive and bipolar disorders”. In: *Biomedical Signal Processing and Control* 72 (2022), p. 103370.
- [61] Zhuangzhuang Li et al. “Deep learning based automatic diagnosis of first-episode psychosis, bipolar disorder and healthy controls”. In: *Computerized Medical Imaging and Graphics* 89 (2021), p. 101882.
- [62] Jia-Yuan Liu et al. “Effects of smartphone-based interventions and monitoring on bipolar disorder: A systematic review and meta-analysis”. In: *World Journal of Psychiatry* 10 (Nov. 2020), pp. 272–285.
- [63] Scott M. Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, 4768–4777.
- [64] Alban Maxhuni et al. “Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients”. In: *Pervasive and Mobile Computing* 31 (2016), pp. 50–66.
- [65] Roger S McIntyre et al. “Bipolar disorders”. In: *The Lancet* 396.10265 (2020), 1841 – 1856.

- [66] Sigurd Melbye et al. “Mood, activity, and sleep measured via daily smartphone-based self-monitoring in young patients with newly diagnosed bipolar disorder, their unaffected relatives and healthy control individuals”. In: *European Child Adolescent Psychiatry* 30 (Aug. 2021).
- [67] Sebastian Mildiner Moraga et al. “Evidence for mood instability in patients with bipolar disorder: Applying multilevel hidden Markov modeling to intensive longitudinal ecological momentary assessment data.” In: *Journal of Psychopathology and Clinical Science* 133.6 (2024), pp. 456–468.
- [68] Amir Muaremi et al. “Assessing bipolar episodes using speech cues derived from phone calls”. In: *International symposium on pervasive computing paradigms for mental health*. Springer. 2014, pp. 103–114.
- [69] James C Mundt et al. “Vocal acoustic biomarkers of depression severity and treatment response”. In: *Biological psychiatry* 72.7 (2012), pp. 580–587.
- [70] Abraham Nunes et al. “Using structural MRI to identify bipolar disorders – 13 site machine learning study in 3020 individuals from the ENIGMA Bipolar Disorders Working Group”. In: *Molecular Psychiatry* 25.9 (Sept. 2020), pp. 2130–2143.
- [71] Venet Osmani et al. “Monitoring activity of patients with bipolar disorder using smart phones”. In: *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*. MoMM ’13. Vienna, Austria: Association for Computing Machinery, 2013, 85–92.
- [72] M. D. Peláez-Aguilera et al. “Fuzzy linguistic protoforms to summarize heart rate streams of patients with ischemic heart disease”. In: *Complexity* (2019).
- [73] Yatharth Ranjan et al. “RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices”. In: *JMIR Mhealth Uhealth* 7.8 (2019), e11734.
- [74] Khansa Rasheed et al. “Explainable, trustworthy, and ethical machine learning for healthcare: A survey”. In: *Computers in Biology and Medicine* 149 (2022), p. 106043.
- [75] Daniele Ravì et al. “Deep Learning for Health Informatics”. In: *IEEE Journal of Biomedical and Health Informatics* 21.1 (2017), pp. 4–21.
- [76] Emna Rejaibi et al. “MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech”. In: *Biomedical Signal Processing and Control* 71 (2022), p. 103107.

- [77] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. “A review of feature selection techniques in bioinformatics”. In: *Bioinformatics* 23.19 (2007), 2507 – 2517.
- [78] Raymond Salvador et al. “Evaluation of machine learning algorithms and structural features for optimal MRI-based diagnostic prediction in psychosis”. In: *PLOS ONE* 12.4 (Apr. 2017), pp. 1–24.
- [79] Marlena Sokół-Szawłowska, Olga Kamińska, and Małgorzata Sochacka. “MoodMon: novel optimization of bipolar disorder monitoring through patient-driven voice parameter submission and AI technology”. In: *Advances in Psychiatry and Neurology/Postępy Psychiatrii i Neurologii* 33.4 (2024), pp. 230–240.
- [80] Oleksandr Sverdlov et al. “A Study of Novel Exploratory Tools, Digital Technologies, and Central Nervous System Biomarkers to Characterize Unipolar Depression”. In: *Frontiers in Psychiatry* 12 (May 2021).
- [81] Marlena Sokol Szawłowska. *Choroba Afektywna dwubiegunowa*. Instytut Psychiatrii i Neurologii, 2013.
- [82] Paweł Teisseyre and Tomasz Klonecki. “Controlling Costs in Feature Selection: Information Theoretic Approach”. In: *Proceedings of the International Conference on Computational Science*. ICCS’21. 2021, pp. 483–496.
- [83] Erico Tjoa and Cuntai Guan. “A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11 (2021), pp. 4793–4813.
- [84] John Torous et al. “The Emerging Imperative for a Consensus Approach Toward the Rating and Clinical Recommendation of Mental Health Apps”. In: *The Journal of Nervous and Mental Disease* 206.8 (2018).
- [85] Maria D. L. A. Vazquez-Montes et al. “Control charts for monitoring mood stability as a predictor of severe episodes in patients with bipolar disorder”. In: *International Journal of Bipolar Disorders* 6.1 (Apr. 2018), p. 7.
- [86] H.U. Wittchen et al. “The size and burden of mental disorders and other disorders of the brain in Europe 2010”. In: *European Neuropsychopharmacology* 21.9 (2011), pp. 655–679.
- [87] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. “Detecting depression severity from vocal prosody”. In: *IEEE transactions on affective computing* 4.2 (2012), pp. 142–150.

- [88] Sana Yasin et al. “EEG based Major Depressive disorder and Bipolar disorder detection using Neural Networks:A review”. In: *Computer Methods and Programs in Biomedicine* 202 (Feb. 2021), p. 106007.
- [89] R. C. Young et al. “A Rating Scale for Mania: Reliability, Validity and Sensitivity”. In: *British Journal of Psychiatry* 133.5 (1978), 429–435.
- [90] Pan Z et al. “Detecting Manic State of Bipolar Disorder Based on Support Vector Machine and Gaussian Mixture Model Using Spontaneous Speech”. In: *Psychiatry investigation* 15 (2018), pp. 695–700.
- [91] John Zulueta et al. “Predicting Mood Disturbance Severity with Mobile Phone Keystroke Metadata: The BiAffect Digital Phenotyping Study (Preprint)”. In: *Journal of Medical Internet Research* 20 (Jan. 2018).

Appendix

Short description of the acoustic parameters extracted from phone calls from the OpenSmile library [23] used in Results section.

Variable Name	Description
pcm.L0Genenergy_sma	Signal energy, logarithmic measure of energy (in dB), also used in separation of voiced and unvoiced speech.
pcm.zcr_sma	Zero-crossing rate (Time Domain Descriptors) - the number of sign changes per unit of time. A parameter for measuring smoothness of a signal, to separate the voiced and unvoiced parts of speech from a speech signal. Zero crossing rates are low for voiced parts and high for unvoiced parts.
voiceprob_sma	Probability of voicing - a measure of voicedness, indicates how close the signal is to an ideal harmonic signal (high probability) or to a noise-like signal (low probability).
f0_sma	Fundamental frequency of a speech signal closely related to pitch, which is our perception of fundamental frequency. Refers to the approximate frequency of the periodic structure of voiced speech signals. The oscillation originates from the vocal folds, which oscillate in the airflow when appropriately tensed. The fundamental frequency is defined as the average number of oscillations per second and expressed in Hertz. It is not exactly periodic but contains significant fluctuations known as jitter and shimmer. F0 is typically not stationary and changes within a sentence, used for expressive purposes like emphasis and questions. Typically ranges from 80 to 450 Hz, with males having lower voices than females and children. The F0 of an individual speaker depends primarily on the length of the vocal folds, correlated with overall body size.
f0env_sma	The last F0 value from the previous voiced region during the adjacent unvoiced region.
pcm.fftMag.fband 0.250_sma	Energy band 250Hz, applied functionalities.
pcm.fftMag.fband 0.650_sma	Energy band 650Hz, applied functionalities.
pcm.fftMag.spectral Roll0ff25_0_sma	Spectral Roll-off point - the frequency in Hz below which a pre-defined percentage of the total energy is concentrated.
pcm.fftMag.spectral Roll0ff50_0_sma	Spectral Roll-off point - the frequency in Hz below which a pre-defined percentage of the total energy is concentrated.
pcm.fftMag.spectral Roll0ff75_0_sma	Spectral Roll-off point - the frequency in Hz below which a pre-defined percentage of the total energy is concentrated.
pcm.fftMag.spectral Roll0ff90_0_sma	Spectral Roll-off point - the frequency in Hz below which a pre-defined percentage of the total energy is concentrated.
pcm.fftMag.spectralflux_sma	Reflects the dynamics of change in the spectral shape and amplitude from one frame to the next. Useful for distinguishing between music and speech signals since music has a higher rate of change.
pcm.fftMag.spectral centroid_sma	Indicates where the center of mass of the spectrum is located, related to the brightness of the sound. It is one of the most frequently-used timbre descriptors.

Variable Name	Description
pcm.fftMag_spectral_maxpos_sma	Spectral maximum position, the position of the maximum value.
pcm.fftMag_spectral_minpos_sma	Spectral minimum position, the position of the minimum value.
f0final_sma jitterlocal_sma	Fundamental frequency – pitch. Measurements of F0 disturbance, reflecting vocal fold vibrations. Describes micro-prosodic variations within short-time frames. Local jitter indicates the deviation in individual F0 period lengths, suggesting irregular closure and asymmetric vocal-fold vibrations. Affected by lack of control of vibration of the cords. Voice qualities include harsh, tense, breathy, creaky voice, and whisper.
jitterddp_sma	The delta period-to-period Jitter, defined as the difference between two consecutive jitter local values.
shimmerlocal_sma	Voice quality feature related to the amplitude variation of the sound wave. Differences in the peak amplitudes of consecutive F0 periods indicate irregularities in voice intensity. Shimmer changes with the reduction of glottal resistance and mass lesions on the vocal cords and is correlated with the presence of noise emission and breathiness.
loghnr_sma	The Harmonics-to-Noise Ratio (HNR) assesses the ratio between periodic components and non-periodic components in a segment of voiced speech. Periodic components arise from vocal cord vibrations, while non-periodic components come from glottal noise, expressed in dB. A higher HNR indicates more efficient speech with greater airflow energy in vocal cord vibrations, associated with sonorant and harmonic voice. A lower HNR denotes asthenic voice and dysphonia.
pcm.rmsenergy_sma	Root Mean Square (RMS) energy.
audSpec_Rfilt_sma_compare_0_to25	Relative Spectral Transform (RASTA auditory bands 1-26). RASTA-style bandpass filter applied to Mel-frequency power-spectral bands raw values. This loudness measure reflects the loudness contained in the speech signal.
pcm.fftMag_fband250.650_sma_compare	Band energy 25–650Hz.
pcm.fftMag_fband1000.4000_sma_compare	Band energy 1-4 kHz.
pcm.fftMag_spectralentropy_sma_compare	Spectral entropy relates to the peakedness of the spectrum, similar to spectral flatness. It captures the peakiness of the spectral representation. The feature is low for spectra with many distinct peaks and high for flat spectra.
pcm.fftMag_spectralvariance_sma_compare	Spectral variance (spectral spread), one of the spectral moments including the second, third, and fourth-order moments.
pcm.fftMag_spectralskewness_sma_compare	Spectral skewness measures the symmetry of the spectrum around its arithmetic mean. The feature is zero for silent segments and high for voiced speech with substantial energy around the fundamental frequency.

Variable Name	Description
pcm_fftmag_spectralkurtosis_sma_compare	Spectral kurtosis measures whether the spectrum can be characterized by a Gaussian distribution. The feature is zero for a Gaussian distribution.
pcm_fftmag_psysharpness_sma_compare / harmonicity	Perceived sharpness/harmonicity of a sound, linked to the spectral characteristics of the sound. A high-frequency signal has a high value of sharpness.
loudness_sma3	Loudness, measures the sound strength.
alphanratio_sma3	Spectral descriptor defined as the ratio between the energy in the low-frequency region and the high-frequency region. Specifically, it is the ratio between the summed energy from 50 to 1000 Hz and 1 to 5 kHz.
hammarbergindex_sma3	Spectral descriptor defined as the ratio of the strongest energy peak in the 0–2 kHz region to that of the strongest peak in the 2–5 kHz region. A fixed static pivot point of 2 kHz separates the low and high-frequency regions.
slope0.500_sma3	Spectral slope as a spectral descriptor reflecting voice quality. Measures the overall shape of a spectrum from 0–500Hz. Can be computed over the full range of the spectrum or over a sub-band.
slope500.1500_sma3	Spectral slope as a spectral descriptor reflecting voice quality. Measures the overall shape of a spectrum from 500-1500Hz. Can be computed over the full range of the spectrum or over a sub-band.
F0semitoneFrom27_5Hz_sma3nz	Fundamental frequency – pitch.
logRelF0.H1.H2_sma3nz	Harmonic ratios computed from the relative amplitudes of F0 harmonics: H1–H2.
logRelF0.H1.A3_sma3nz	The ratio of the first harmonic (H1) to the highest harmonic in the third formant range (A3). The amplitudes of the F0 harmonics contain valuable information about the state of the vocal tract.
f1/2/3frequency_sma3nz, f1/2/3bandwidth_sma3nz, f1/2/3amplitudelogrelf0_sma3nz	Formants - resonance frequencies (peaks) of the vocal tract system used to uniquely identify vowels. The fundamental frequency is denoted by F0, and formants are higher-order resonance frequencies excited by F0. F1, F2, F3 measure speech formant frequency (in Hz), speech formant bandwidth (in Hz) of F0, and amplitude.
pcm_fftMag_mfcc_0_ to pcm_fftMag_mfcc_12_	Mel-Frequency Cepstral Coefficients (MFCC), composed of 13 cepstral descriptors, representing the short-term power spectrum of a sound. MFCCs are the linear cosine transform of a short-term log power spectrum on a non-linear Mel scale of frequency. Used in applications including speaker verification, speaker recognition, and emotion detection.