Wearable Wellness: Depression Screening via Fitbit Data Collected During COVID-19 Pandemic

Nikola Grozdani
Computer Science
Worcester Polytechnic Institute
Worcester, MA, USA
ngrozdani@wpi.edu

America Muñoz
Computer Science/Information Systems
National Louis University
Chicago, IL, USA
amunoz8@my.nl.edu

Alexander Pietrick
Computer Science/Data Science
Worcester Polytechnic Institute
Worcester, MA, USA
ajpietrick@wpi.edu

Ricardo Flores
Data Science
Worcester Polytechnic Institute
Worcester, MA, USA
rflores@wpi.edu
ORCID: 0000-0002-9466-3993

Avantika Shrestha
Data Science
Worcester Polytechnic Institute
Worcester, MA, USA
ashrestha4@wpi.edu
ORCID: 0000-0002-7263-100X

Xingtong Guo
Architectural Engineering
Worcester Polytechnic Institute
Worcester, MA, USA
xguo3@wpi.edu
ORCID: 0000-0002-2984-7841

Shichao Liu
Architectural Engineering
Worcester Polytechnic Institute
Worcester, MA, USA
sliu8@wpi.edu
ORCID: 0000-0002-0642-9700

Elke A. Rundensteiner
Computer Science/Data Science
Worcester Polytechnic Institute
Worcester, MA, USA
rundenst@wpi.edu
ORCID: 0000-0001-5375-9254

***Abstract*—The COVID-19 pandemic has greatly increased depression among adolescents. The current depression diagnosis process requires significant patient effort and can be costly. Prior research through passively collected data has shown promising depression screening results but is limited by complex data collection and privacy concerns. In this research, we create multiple machine learning models to screen physiological data collected from Fitbit, a wearable biomarker, and depression screening surveys across 166 college students. The highest-scoring model on these physiological modalities achieved an F1-score of 0.92. Our research findings highlight the potential impact of digital technology development in current clinical practices.**

Index Terms*—*passive depression screening, physiological data, time series, wearable biomarker, COVID-19

I. Introduction

Depression is one of the most common mental health disorders in the US. It is particularly common among college students between $18-25$ years, with 44% experiencing depressive symptoms in $2021-2022$ [1]. The COVID-19 outbreak contributed to the prevalence of depression due to the lack of socialization and time spent indoors [2]. In 2021, 32.8% of adults experienced an increase in depression symptoms compared to 27.8% in the earlier months of the COVID-19 pandemic [3]. Early depression diagnosis and treatment is crucial as it is one of the most frequent mental illnesses associated with suicide [4]. The current process of depression diagnosis includes screening surveys deployed by mental health professionals and lengthy clinical interviews [5]. Diagnosis requires significant patient effort and can be costly [6]; it also heavily relies on the ability to recognize one's symptoms and seek treatment [7]. Therefore, the application of machine learning algorithms to test passively collected data from digital technology has become more prevalent. Prior research using passively collected data includes text [7], [8], audio [9]–[11], facial features [5], and video screening [12], [13] have shown promising depression screening results but are limited by privacy concerns and complex data collection.

The use of digital biomarkers to passively monitor heart rate, sleep, and motion can prove valuable insights into mental health research and current clinical practices [14]–[16]. Digital bio-markers can monitor a person's physiological data daily they also prove to be a passive form of data collection. Additionally, digital bio-markers have a widespread reach with 1 in 5 Americans owning a smartwatch in 2020 [17]. These wearable biomarkers can track a variety of physiological modalities including activity time, movement speed, and step counts [16]. Moreover, these physiological modalities were found to be directly correlated with depression [16]. As such we utilize data collected from Fitbit, a wearable biomarker, to analyze the correlation between physiological activity and depression survey scores.

The time series data collected from 166 college students consists of five physiological modalities (steps taken, distance traveled, calories burned, heart rate, and sleep). Previous

979-8-3503-0965-2/23/$31.00 ©2023 IEEE

Fig. 1. Current Depression Screening Process

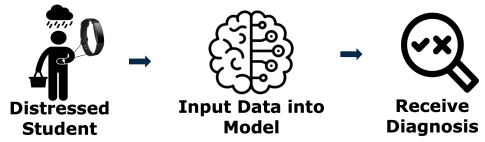

Fig. 2. Passive Depression Screening Process

research on passively collected data has used deep learning methods to screen for depression [10], [12], [13], [16]. Although promising results can be found, a deep learning approach is a computationally expensive method that is predominately used for larger datasets and produces results that could be difficult to interpret. Therefore in this research, we use the passively collected data on 5 different machine learning models. The time series data is input into the models with a variety of semantic sampling and modality combinations. Through the use of the Time Series Feature Extraction Library, a total of 177 features are extracted from the time sensor data to determine what modalities are most present in the highest-scoring models. Our contributions include:

1) The analysis of a unique dataset observing the passively collected physiological data of 166 adolescents, during the COVID-19 pandemic.
2) Transforming the time series dataset consisting of 100,000 hours into a format that is compatible with a machine learning model.
3) Through different sampling frequencies, we were able to comprehend distinct periods of activities in correlation to their depression scores.

II. Data

Beginning in Summer 2020 through Spring 2021, 166 students enrolled in Worcester Polytechnic Institute (WPI), participated in a variety of data collection simultaneously. The students were monitored in 7 cohorts for the periods mentioned above, ranging from 3 to 36 students in each cohort.

Students were provided with a Fitbit Inspire HR, to be worn every day throughout the data collection. The Fitbit monitored calories burned, distance traveled, steps taken, sleep patterns, and heart rate. All modalities were tracked in minutes with the exception of heart rate which was tracked in seconds.

In addition to that, students were directed to complete the Center for Epidemiologic Studies Depression Scale Revised 10 (CESD-R10) surveys monthly [18]. The CESD-R10 survey is a 10-question self-report survey that inquires on how a person may have felt or behaved and is commonly used in research settings. It has proven to show valid results and is considered for use among adults and adolescents [19]. The sum of the total score of 10 or above is indicative of depressive symptoms. In this research, all Fitbit data collected and CESD-R10 survey results were utilized. However, only 99 students were included in our study as they had data from both Fitbit and the survey.

A. Descriptive Analytics

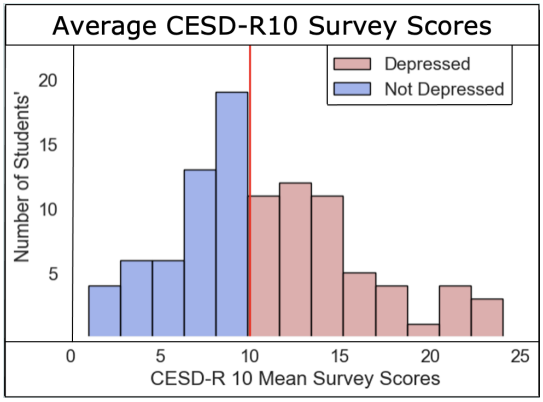

Fig. 3. Average CESD-R10 Scores for All Cohorts

Fig 3 is a bar chart that illustrates the distribution of the 166 students in the dataset based on their CESD-R10 mean survey scores. Here, it is shown that 51% of students are labeled as "Depressed" indicating their survey score was greater or equal to 10. Whereas the other 49% of students scored below the threshold and were labeled as "Not Depressed." This distribution highlights a relatively balanced representation of depressed versus non-depressed students within our dataset.

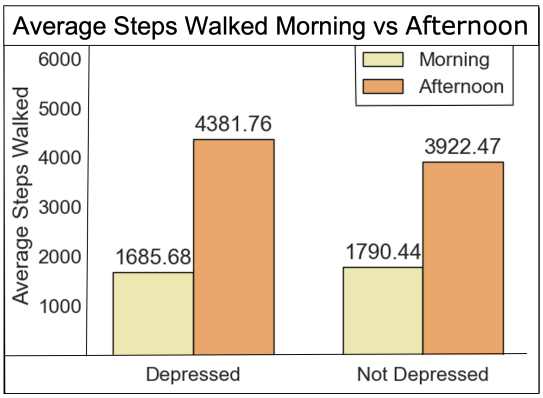

Fig. 4. Average Steps: Morning vs Afternoon Based on CESD-R10 Label

Fig. 4 visualizes the unequivocal differences between data recorded in the morning and data recorded in the afternoon. As the average steps taken by both depressed and not depressed during morning hours $[00:00-12:00]$ was nearly 1700 steps. Meanwhile, the average number of steps taken during afternoon hours $[12:00-00:00]$ is closer to 4000 steps.

III. Methodology

A. Data Preprocessing

This data prepossessing process encompassed various essential tasks, including data concatenation, aggregation, and imputation of missing values, each step vital for our research.

1) Data Concatenation: Our first step featured data concatenation on each of the individual modalities in the dataset. This process involved taking all of a student's time series data across a modality and consolidating it into one file. Simultaneously, we aligned each student's modality data with their corresponding depression scores. This merging process resulted in a centralized dataset where potential issues related to data mismatches were minimized. By unifying modalities for each student and their respective depression scores, we allow for more holistic insights into relations within our data.

2) Aggregation: After data concatenation, we proceeded with the aggregation process, which involved creating daily representations of the students' modalities. To achieve these values, we calculated the mean for heart rate and the sum for calories burned, steps taken, distance traveled, and hours slept.

For heart rate, the mean value provided insights into each student's typical daily level of exertion, aligning with established clinical practices for assessing cardiovascular health [20]. On the other hand, for calories burned, steps taken, distance traveled, and hours slept, the summing process allowed us to capture the total daily values for each of these modalities. These values provided a comprehensive representation of the students' common health and fitness practices [21].

3) Imputation: Throughout the various steps in our preprocessing pipeline, missing values were encountered. These missing values occurred on days when students did not have any recorded Fitbit data. To maintain consistency, most students had at least 85 days of data for each of the different modalities; this became the cutoff point for usable data segments due to availability constraints. For students who had more than 85 days of data, we selected the first 85 days to maintain consistency and compatibility across the dataset. For students who had slightly less than 85 days of data or had entire days of missing data, we addressed the missing values by employing a mean imputation method. Mean imputation is a common and straightforward technique that replaces missing values with the mean value of the corresponding modality, thereby preserving the overall central tendency of the data [22]. The value was generated by only considering the cohort of which the student had a missing value and then taking the mean of the modality in question.

B. Sampling Frequency

While biomarkers are designed to constantly record data, not all data collected throughout the day may be crucial in depression screening. Through the data exploration process, it was discovered that there was a distinct difference between physiological data collected during morning hours and physiological data collected during afternoon hours. The distinctions provided a new direction of input to be tested in the modeling sequence. This direction may be critical in applications as knowing when to collect data could help streamline the passive depression screening. Additionally, different samplings across the modalities throughout periods of the day could potentially reveal insightful trends across all of the data.

We consider these factors by employing two main sample frequencies: daily (24-hour basis) and 12-hour basis. This approach provided a detailed view of the students' fitness patterns during distinct periods of the day (full day, morning, and afternoon) and allowed us to comparatively draw conclusions on which intervals of the day are most important for screening.

C. Modeling Sequence

After the time series data was transformed into a more consistent and uniform matter, it was then used to produce models for passive depression screening.

1) Data Sampling: Modeling was conducted over four distinct samplings of data. Each of these was run through the same modeling independent of each other. These semantic samplings were based on how the data was divided; predominantly on whether the modalities were split by 24 hours or split by 12 hours. The 12 hour samplings can be further subdivided into two groups where the morning consisted of data between hours $00:00$ and $12:00$ and the afternoon consisted of data between hours $12:00$ and $00:00$. The fourth sampling was constructed through a concatenation of both the morning and afternoon 12 hours. These will be referred to as the 24 Hour, 12 Hour: Morning, 12 Hour: Afternoon, and 12 Hour: Morning + Afternoon samplings going forward.

For each of the data samplings, different modalities were used as input to the modeling sequence. For the 24 hour, 12 Hour: Morning, and 12 Hour: Afternoon samplings, the 5 independent modalities were constructed according to the previously stated specifications. For the 12 Hour: Morning + Afternoon sampling, 10 files were used (all files across the 12 Hour: Morning and 12 Hour: Afternoon samplings). Within each sampling, all different combinations were tested as inputs for the modeling sequence. For samplings with 5 files, this led to 32 unique modality combinations, and with the sampling of 10 files, this led to 1023 unique modality combinations.

2) Feature Extraction from Time Series: For each modality within a combination, its raw time series was transformed into a list of features using the well-known Python package Time Series Feature Extraction Library (TSFEL) [23]. TSFEL emerged as the optimal choice due to its ability to transform time series data into rich features with minimal computational cost. The library offers an array of distinct features, encompassing spectral, temporal, and statistical domains. The features within these domains aim to capture information covering frequency-based attributes, temporal patterns, and statistical metrics. All possible features offered by the TSFEL library were utilized, leading to a total of 177 distinct features extracted from a single time series. The feature arrays produced on different modalities within a combination were then stitched together to form a comprehensive list of features.

This list of features served as continuous combinations of the modalities through the modeling sequence.

3) Dimension Reduction: A dimension reduction was then conducted to scale values within a reasonable computational range and to remove features that provided no unique insights into the data's variability. This was completed through the use of Sklearn's [24] MinMaxScaler and PCA [25] respectively.

4) Modeling: An array of different models were used in order to test a wide range of machine-learning techniques. These models included Logistical Regression, Binary Classification, Support Vector Machine (SVM), and Random Forest Classifier (RFC) [26]. Sklearn's [24] implementations with default hyper-parameters were used for all of these models with the exception of Binary Classification, where XGBoost was utilized. An $80/20$ training/testing split was used for all models. In total, $5,595$ models were trained across the $1,119$ unique modality combinations.

IV. Results

A. Metric Evaluation

For each model, an array of different metrics were used to evaluate different aspects of the model. For comparing models across the different data samplings, we had to decide on a singular metric to use. While accuracy is the most common metric for classification tasks, the $F1$ score is preferred in healthcare because of its higher emphasis on instances of true positives. Additionally, the $F1$ score lets us easily compare with similar studies [9], [12] Therefore, we use the $F1$ score as defined in equation (1), as the metric to evaluate our models.

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{1}$$

where TP is the number of true positives, FP the number of false positives, and FN the number of false negatives.

B. Across Data Samplings

Fig. 5 highlights the different time sampling methods used and their respective F1 scores across the 5 models. All samplings have an F1 score above 0.70 across all models. Moreover, the Random Classier Model features the 12 Hour: Afternoon time sampling which produced the second highest F1 score at 0.91. The highest scoring features consist of the concatenated Morning and Afternoon at a 0.92 F1 score.

C. Across Data Modalities

The plot in Fig. 6 shows the Fitbit modalities based on the data sampling. Heart Rate had the single highest F1 score at 0.83, indicative of it being the most predictive singular physiological modality for the afternoon sampling. Although, Heart Rate proved to be the single highest modality, a combination of modalities across the 12 Hour: Afternoon and 12 Hour: Morning + Afternoon sampling, scored the highest in F1 score with Afternoon scoring a 0.91 and the Morning + Afternoon sampling at an F1 score of 0.92.

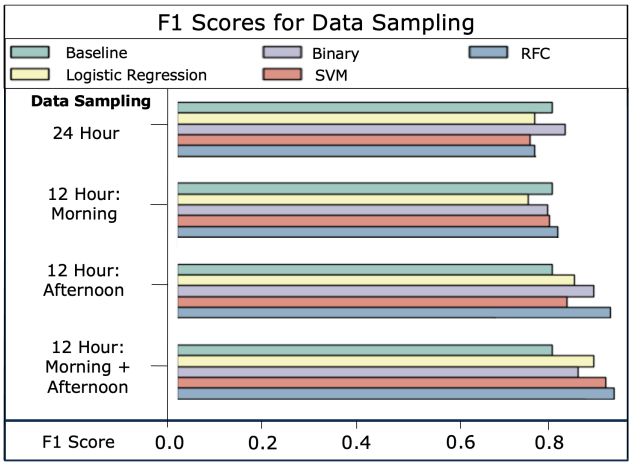

Fig. 5. F1 Scores for Data Sampling

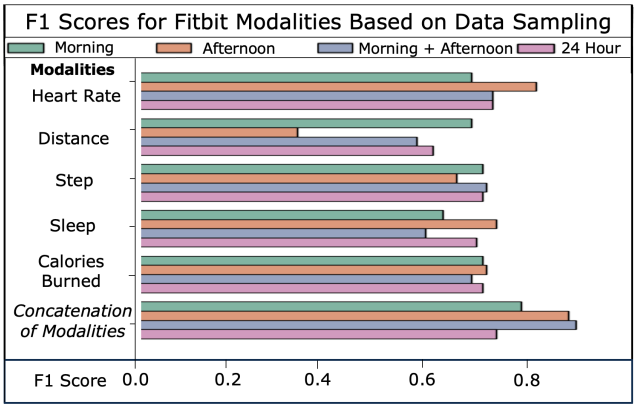

Fig. 6. F1 Scores for Fitbit Modalities Based on Data Sampling

V. Discussion

The combination of morning and afternoon data yields the most robust and consistent outcomes in determining an individual's depressive state. Potential reasons for the distinct differences in morning and afternoon data could be attributed to the dynamic nature of human physiology. Hypothetically, depressed students may wake up later and thus, burn fewer calories or take fewer steps in the morning hours. The data collected during afternoon hours $12:00$ to $00:00$ includes the students' physiological data when they are assumed to be fully awake. Thus, there may be stronger signals in this dataset for those who have screened positive for depression.

Among the various modalities examined, heart rate emerged as the most influential feature in depression screening. Prior research has established a strong connection between the variability of heart rate and depression [27]. However, the combination of physiological data can provide a more holistic view of an individual's depressive status. The combination of modalities allows for the features to be extracted across multiple high-scoring modalities simultaneously. Specifically, the combination of modalities such as calories, steps, distance, and heart rate has proven to have strong screening results.

A. Limitations and Future Work

Although the results prove to show a relationship between physiological data and depression screening; we believe they can be improved as missing values were present in our dataset. There were various periods of time, where Fitbit data was not collected due to user-induced hardware inactivity. Similarly, there were instances where Fitbit data was collected but survey scores were not. A decrease in the rate of missing data would provide models with more complete information to utilize.

In this experiment, we applied the mean imputation approach in instances where missing values were found. Although the mean imputation method is a widely accepted method, there are several alternative approaches that could be used to preserve the integrity and account for any uncertainties in the data. For instance, techniques such as forward fill, backward fill, interpolation, or using machine learning algorithms for imputation can be leveraged.

While our process of aggregation used mean and sum methods, it is important to note that there are other aggregation methods such as median, maximum, or percentile that could offer different perspectives on the students' fitness metrics [28]. Future testing with sequential transfer models could provide more substantial results than those from only classification models. These models have the capability to capture complex patterns in time series data that might not be found by a classification model on features extracted from time series.

Finally, while the combination of daily and 12-hour basis sampling frequencies provided sufficient data for our analysis, it is also worth noting that exploring other sampling frequencies, such as weekly or monthly basis, could have enhanced the flexibility and applicability of the study even further.

VI. Conclusion

Our study provides compelling evidence that physiological data has significant potential for depression screening. Combining morning and afternoon data, along with multiple physiological modalities resulted in the successful training of a machine learning model, ultimately producing an F1 score of 0.92. As technology continues to advance and the availability of wearable devices increases, we believe the use of physiological data for depression screening holds promising opportunities for future research and clinical applications.

Acknowledgments

The authors would like to thank the National Science Foundation for funding this research under the following grants $\#1852498$, $\#2103832$, $\#2028224$, and $\#1910880$. We would also like to thank Prof. Rodriguez, Prof. Liu, Guo, Wang, and Noggle for data collection.

References

[1] S. K. Lipson, S. Zhou *et al.*, "Trends in college student mental health and help-seeking by race/ethnicity: Findings from the national healthy minds study, 2013–2021," *Journal of Affective Disorders*, vol. 306, pp. 138–147, 2022.

[2] V. J. Clemente-Suárez, M. B. Martínez-González *et al.*, "The impact of the covid-19 pandemic on mental disorders. a critical review," *International Journal of Environmental Research and Public Health*, vol. 18, no. 19, p. 10041, 2021.

[3] C. K. Ettman, G. H. Cohen *et al.*, "Persistent depressive symptoms during covid-19: a national, population-representative, longitudinal study of us adults," *The Lancet Regional Health–Americas*, vol. 5, 2022.

[4] J. M. Bertolote and A. Fleischmann, "Suicide and psychiatric diagnosis: a worldwide perspective," *World psychiatry*, vol. 1, no. 3, p. 181, 2002.

[5] R. Flores, M. Tlachac *et al.*, "Temporal facial features for depression screening," in *UbiComp/ISWC '22 Adjunct*, 2022, pp. 488–493.

[6] J. Lu, C. Shang *et al.*, "Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning," *IMWUT '18*, vol. 2, no. 1, pp. 1–21, 2018.

[7] A. Shrestha, M. Tlachac *et al.*, "Bert variants for depression screening with typed and transcribed responses," in *UbiComp/ISWC '22 Adjunct*, 2022, pp. 211–215.

[8] M. Tlachac and S. S. Ogden, "Left on read: Reply latency for anxiety & depression screening," in *UbiComp/ISWC '22 Adjunct*, ser. UbiComp/ISWC '22 Adjunct. ACM, 2023, p. 500–502.

[9] E. Toto, M. Tlachac, and E. A. Rundensteiner, "Audibert: A deep transfer learning multimodal classification framework for depression screening," in *Proceedings of the 30th ACM international conference on information & knowledge management*, 2021, pp. 4145–4154.

[10] R. Flores, M. Tlachac *et al.*, "Depression screening using deep learning on follow-up questions in clinical interviews," in *2021 20th IEEE ICMLA)*. IEEE, 2021, pp. 595–600.

[11] X. Ma, H. Yang *et al.*, "Depaudionet: An efficient deep model for audio based depression classification," in *AVEC '16*, 2016, pp. 35–42.

[12] R. Flores, M. Tlachac *et al.*, "Audiface: Multimodal deep learning for depression screening," in *Machine Learning for Healthcare Conference*. PMLR, 2022, pp. 609–630.

[13] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognition Letters*, vol. 146, pp. 1–7, 2021.

[14] V. De Angel, S. Lewis *et al.*, "Digital health tools for the passive monitoring of depression: a systematic review of methods," *NPJ digital medicine*, vol. 5, no. 1, p. 3, 2022.

[15] N. C. Jacobson and Y. J. Chung, "Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones," *Sensors*, vol. 20, no. 12, p. 3572, 2020.

[16] K. Watanabe and A. Tsutsumi, "The passive monitoring of depression and anxiety among workers using digital biomarkers based on their physical activity and working conditions: 2-week longitudinal study," *JMIR Formative Research*, vol. 6, no. 11, p. e40339, 2022.

[17] E. A. Vogels, "About one-in-five americans use a smart watch or fitness tracker," 2020.

[18] E. M. Andresen, K. Byers *et al.*, "Performance of the 10-item center for epidemiologic studies depression scale for caregiving research," *SAGE Open Medicine*, vol. 1, p. 2050312113514576, 2013.

[19] E. E. Haroz, M. L. Ybarra, and W. W. Eaton, "Psychometric evaluation of a self-report scale to measure adolescent depression: the cesdr-10 in two national adolescent samples in the united states," *Journal of affective disorders*, vol. 158, pp. 154–160, 2014.

[20] B. Olshansky, F. Ricci, and A. Fedorowski, "Importance of resting heart rate," *Trends in Cardiovascular Medicine*, 2022.

[21] B. Elsawy and K. E. Higgins, "Physical activity guidelines for older adults," *American family physician*, vol. 81, no. 1, pp. 55–59, 2010.

[22] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine*, vol. 4, no. 1, 2016.

[23] M. Barandas, D. Folgado *et al.*, "Tsfel: Time series feature extraction library," *SoftwareX*, vol. 11, p. 100456, 2020.

[24] O. Kramer and O. Kramer, "Scikit-learn," *Machine learning for evolution strategies*, pp. 45–53, 2016.

[25] P. Geladi and J. Linderholm, "Principal component analysis," 2020.

[26] G. James, D. Witten *et al.*, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[27] M. Čukić, D. Savić *et al.*, "When heart beats differently in depression: Review of nonlinear heart rate variability measures," *JMIR Mental Health*, vol. 10, no. 1, p. e40342, 2023.

[28] A. Almeida, P. de Villiers *et al.*, "Visual comparison of statistical feature aggregation methods for video-based similarity applications," in *2020 IEEE FUSION*. IEEE, 2020, pp. 1–8.